PRENATAL RISK ASSESSMENT OF DOWN SYNDROME BY PROBABILISTIC CLASSIFIERS

by

Ömer Uzun

B.S., Computer Engineering, Işık University, 2007

Submitted to the Institute for Graduate Studies in

Science and Engineering in partial fulfilment of

the requirements for the degree of

Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2013

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis supervisor Prof. Fikret Gürgen, for his support throughout this study with not only the subject itself but also many other common interests of ours. With his significant contribution, I was able to finish this thesis work.

I am grateful to present my thanks to Prof. Füsun Varol, Özlem Özer, who shares Down syndrome dataset to use in this thesis work.

I am thankful to Prof. Elizabeth Thom, Julia Zachary and Mark McNellis from George Washington University Biostatistics Center who also shares Down syndrome dataset to use in this thesis work.

I would like to thank my friend Mert for his guidance, support during my study and acting as a mentor to me.

I cannot forget the support of my family, with their existence and pure love. The most special thanks goes to my best partner and friend, my wife. Esma, you gave me your unconditional support and love through all this long process.

ABSTRACT

PRENATAL RISK ASSESSMENT OF DOWN SYNDROME BY PROBABILISTIC CLASSIFIERS

Over the last 20 years, new technology has improved the methods of detection of fetal abnormalities, including Down syndrome. While there are ways to diagnose Down syndrome by obtaining fetal tissue samples by amniocentesis or chorionic villus sampling, it would not be appropriate to examine every pregnancy this way. Besides greatly increasing the cost of medical care, these methods do carry a slight amount of risk to the fetus. So non-invasive methods such as characteristics and screening analysis have been developed to try to identify those pregnancies at "high risk". These pregnancies are then candidates for further diagnostic testing. In this thesis, we address the decision-making problems in diagnosing Down syndrome cases from the machine learning perspective aiming to decrease invasive tests. Initially, we present a comprehensive and comparative analysis of the classification techniques in Down syndrome prediction. In parallel, we evaluate the predictor effects of input features in order to eliminate the redundant features and decide the optimum feature subset leading to the highest prediction performance. Later, we focus on improving the classification performance either by parameter optimization or by improving the information content of the data. First we handle the problem of imbalanced class distribution. As a solution to imbalance class problem we analyse decision threshold optimization and re-sampling the training data techniques. Secondly, we use probabilistic classifiers based on applying Bayes Theorem, Naive Bayes and Bayesian Networks, to predict the Trisomy 21 case. In contrast to probabilistic classifiers we also apply some of widely used and well known classifiers such as Decision Tree, Support Vector Machine, Multi Layer Perceptron, and k-NN. In this thesis, we aim to evaluate the probabilistic classifiers performance with respect to these methods. This comparison is based on performance metrics such as sensitivity, specificity, accuracy and Receiver Operating Characteristics. The results of the experiments show that (i) probabilistic classifiers enable acceptable prediction of Trisomy 21 case and (ii) the classification performance can be improved by using the proposed techniques in this study.

ÖZET

OLASILIKSAL SINIFLANDIRICILAR İLE DOWN SENDROMUNUN DOĞUM ÖNCESİ RİSKİNİN HESAPLANMASI

Teknolojinin gelişimi ile birlikte Down sendromu gibi genetik düzensizliklerin gebelik sırasında tanımlama yöntemleri oldukça gelişmiştir. Down sendromu teşhisi için fetal doku örneklerini analiz eden amniyosentez veya koryon villus örneklemesi gibi kesin tanı koyan yollar vardır, fakat her hamilelikte bu invaziv yöntemleri kullanmak uygun değildir. Bu yöntemler büyük ölçüde tıbbi bakım maliyetini artırmanın yanı sıra, fetus için risk teşkil etmektedir. Bu sebeple öznitelik ve görüntüleme analizleri gibi invaziv olmayan yöntemler ile bu gebeler "yüksek risk" grubunda sınıflandırılabilmektedir. Bu sınıftaki gebeler daha fazla tanısal test ile değerlendirilmektedir. Bu tezde, invaziv testleri azaltmak amacıyla Down sendromu yüksek riskli sınıfını oluşturmak için karar verme problemleri yapay öğrenme bakış açısı ile ele alınmıştır. İlk olarak, Down Sendromu tahmini için sınıflandırma tekniklerinin kapsamlı bir analizi sunulmuştur. Aynı zamanda, özniteliklerin belirleyici etkileri değerlendirilmiş ve gereksiz değişkenler elenerek ideal öznitelik alt kümesi belirlenmiştir. Çalışmanın devamında, metodolojik iyileştirmeler ve kullanılan veri kümesinin bilgi içeriğinin genişletilmesi ile tahmin performansı arttırılmıştır. İlk olarak, dengesiz sınıf dağılımı problemi ele alınmış, karar eşiği optimizasyonu ve öğrenme kümesinin yeniden örneklenmesi ile çözüm yöntemleri analiz edilmiştir. İkinci olarak, kategorik özniteliklerin sayısal değerlere dönüştürülmesinin tahmin gücüne olan etkisi incelenmiştir. Bu çalışmanın kapsamında iki farklı veri seti kullanılmıştır. Son olarak, Trizomi 21 tahminlemesi için Bayes Teoremini kullanan olasılıksal sınıflandırıcılardan Naive Bayes ve Bayes Ağlar yöntemleri uygulanmıştır. Ayrıca yaygın olarak kullanılan sınıflandırıcılardan Karar ağacı, Destek Vektör Makinesi, Çok Katmanlı İdrak, ve k-NN kullanılmıştır. Ana motivasyonlarımızdan biri olan olasılıksal sınıflandırıcılar ile diğer sınıflandırıcıların performansı duyarlılık, özgüllük, doğruluk ve ROC değerleri esas alınarak karşılaştırılmıştır. Deneylerde (i) olasılıksal sınıflandırıcıların Trizomi 21 tahmininde kabul edilebilir başarı oranı elde ettiği ve (ii) bu çalışmada önerilen teknikler kullanılarak tahmin performansının arttırılabileceği görülmüştür.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS / ABBREVIATIONS

AFP	Alpha fetoprotein
ANN	Artificial neural network
AUC	Area under the curve
BN	Bayesian networks
CRL	Crown rump length
DFE	Discriminative frequency estimate
DR	Detection rate
DS	Down syndrome
DT	Decision Tree
DV	Ductus venosus
f β-hCG	free β-human chorionic gonadotropin
FA	Factor Analysis
FE	Frequency estimates
FN	False negative
FP	False positive
FPR	False positive rate
hCG	Human chorionic gonadotropin
IQ	Intelligence quotient
k-NN	k-Nearest neighbor
MLP	Multi layer perceptron
MoM	Multiples of median

NB	Naive Bayes
NT	Nuchal translucency
PAPP-A	Pregnancy associated plasma protein A
PCA	Principal component analysis, a vector space transform often
RFE	Recursive feature elimination
ROC	Receiver operating characteristic
SLA	Simple learning algorithm
SVM	Support vector machines
T21	Trisomy 21
TAN	Tree augmented network
TN	True negative
TP	True positive
TPR	True positive rate
TR	Tricuspid regurgitation
uE3	Unconjugated estriol

1. INTRODUCTION

Down syndrome also known as Trisomy 21, is a chromosomal condition caused by the presence of all or part of an extra 21st chromosome. It is named after John Langdon Down, the British physician who described the syndrome in 1866. The condition was clinically described earlier in the 19th century by Jean Etienne Dominique Esquirol in 1838 and Edouard Seguin in 1844 [1]. Down syndrome was identified as a chromosome 21 Trisomy by Dr. Jérôme Lejeune in 1959. Down syndrome in a fetus can be identified through chorionic villus sampling or amniocentesis during pregnancy, or in a baby at birth.

The incidence of Down syndrome at birth is approximately 1 in 750. However, since the majority of Trisomy 21 pregnancies spontaneously miscarry, the incidence at conception must be higher, perhaps as high as 1 in 150. The chance of a Trisomy 21 conception rises with advancing maternal age. Thus, in affluent countries where there is a population-based, antenatal screening program for Trisomy 21 and a trend for women to postpone having children until the fourth decade, the number of Down syndrome diagnoses has increased.

Infants who are affected by Down syndrome are usually diagnosed very soon after birth because they have reduced body tone in combination with minor features including flat occiput, upslanting palpebral fissures, epicanthic folds, large or slightly protruding tongue, single palmar crease, small fifth finger, and wide gap between first and second toes. More importantly, these infants also have an increased chance of being affected by one or several different serious congenital malformations or illnesses. Thus, about one in five affected children die before age 5 years and two in five are affected by conditions such as congenital heart defect, bowel atresia, or leukemia. For most but not all families of an affected child, cognitive impairment is the most important complication of the syndrome. This is always present, although of variable severity. In general, the type of cognitive impairment is not specific to Trisomy 21. Delay in development is often evident from early infancy and when IQ is measured, scores indicate moderate to severe retardation (IQ range

10–70). Thus, Down syndrome individuals achieve variable levels of independence in adult life but only a minority are fully independent in all daily living skills.

A major research theme is the identification of pregnancies where the fetus is affected by Trisomy 21. To accomplish this, antenatal screening programs assess independent risk factors such as the mother's age, the maternal serum levels of certain pregnancy-related proteins, and the appearance of the fetus on ultrasound examination. Up to 80% of affected pregnancies may be diagnosed before 20 weeks gestation. Safer diagnostic tests such as chromosome analysis of fetal cells in maternal blood may replace current tests such as amniocentesis and chorionic villus sampling. Antenatal screening for Trisomy 21 raises ethical issues and heated debate, but the argument for decreasing Down syndrome associated health problems is unopposed.

Prenatal screening for Down syndrome was developed by the introduction of nuchal translucency (NT) and ultrasound to the first trimester of pregnancy. In pregnancies with fetal Trisomy 21, low maternal serum pregnancy associated plasma protein A (PAPP-A) and elevated free β-human chorionic gonadotropin (f β-HCG) values were observed by the 1990s [2,3]. Screening for Trisomy 21 by combining maternal age, fetal NT thickness and maternal serum f β-HCG and PAPP-A at 11-13 weeks was associated with a detection rate of about 90% for a false-positive rate of 5% [4, 5]. However, since measurements of NT varied considerably between centers and clinicians, the sensitivity can be as low as 31%, thus it could hardly be reliably incorporated into the test [6].

The prenatal risk assessment of Down syndrome can be modelled using machine learning methods providing automated decision support to clinicians when necessary. On the contrary to the emergence and importance of decision support systems in Down syndrome risk assessment process, the related literature is limited. Statistics analysis models are used as risk estimation in Down syndrome. However, there is not enough study that evaluate the prenatal risk with machine learning methodologies.

1.1. Research Overview

During this thesis, we mainly concentrated on predictive modelling of Down syndrome which may be defined as non-invasive approach. After analyzing the existing statistical models, we performed experiments to build up novel decision support systems as a benchmark study aiming to pave the way for further studies.

Our first research interest is prenatal risk assessment of Down syndrome by probabilistic classifiers including the following subtask: gathering of different datasets, applying of state of the art classifiers comparatively, handling the constraints of the standard methods in order to improve the prediction performance and investigating the effect of the physicians experience as a human factor in success of Down syndrome prediction.

1.1.1. Dataset Gathering and Classifier Selection

The initial step of this thesis is gathering data and choosing acceptable classification techniques for prediction problem. The most challenging problem of machine learning studies in medical domain are related to the data retrieval. Unfortunately, there are no public Down syndrome datasets to be used in machine learning experiments. We contact more than fifty specialists from different universities and can only get positive feedback from two universities to get Down syndrome datasets. Trakya University and George Washington University give consent us to use their Down syndrome datasets within this study. We have used the most popular representatives of different classifier categories because comparative analysis of diverse classifiers enables determination of the best fitting models in application domain. We have used ROC analysis for comparison and evaluation of classification performance. We have performed feature selection and feature extraction in order to reduce the computational cost and improve performance in the rest of the experiments.

1.1.2. Handling the Imbalanced Class Distribution

Both the databases represents an imbalanced distribution of class samples (more than 99% negative class). Sampling methods such as over-sampling and under-sampling have been used to balance the number of instances in the classes. Moreover, we use threshold optimization technique as another solution to imbalance class distribution problem.

1.1.3. Transformation of Nominal Variables Into Numeric Data

Datasets we analyzed include both categorical and numerical values. Transformation of categorical variables into numeric attributes is an important pre-processing stage for distance based algorithms such as Support Vector Machines (SVM), k-Nearest Neighbor (kNN) etc. affecting the performance of the classification. We have used two different techniques to transform categorical variables, frequency based encoding and binary encoding methods.

1.1.4. Analysis for the Effect of Physician Factor

We have analyzed the effect of the experience level of individual physicians in success of detection rate of Down syndrome by comparing the model with two different datasets including same type of features.

This research is mainly concentrated on predictive modelling of Down syndrome procedure as a novel application domain in machine learning community. The proposed modifications to standard machine learning algorithms produce enhanced prediction performance in Down syndrome domain as well as presenting potential of generalization to other real world applications.

1.2. Prenatal Diagnostic Methods

1.2.1. Maternal Age

The incidence of fetal Trisomies is directly related to maternal age [7]. The risk of having a child with Down syndrome increases in a gradual, linear fashion until about age 30 and increases exponentially thereafter as shown in the Figure 1.1 [8]. The risk of having a child with Down syndrome is 1/1300 for a 25 years old woman; at age 35, the risk increases to 1/365. At age 45, the risk of a having a child with Down syndrome increases to 1/30.

Historically, maternal age can be viewed as the first "screening test" for fetal chromosome abnormalities. In the late 1970s, about 5 percent of pregnancies in the United States occurred in women who were 35 years or older [9]. At age 35, the second-trimester prevalence of Trisomy 21 (1/270) approaches the estimated risk of fetal loss due to amniocentesis (1/200) [10]. Therefore, age 35 was chosen as the screening cut-off, the risk threshold at which diagnostic testing is offered.

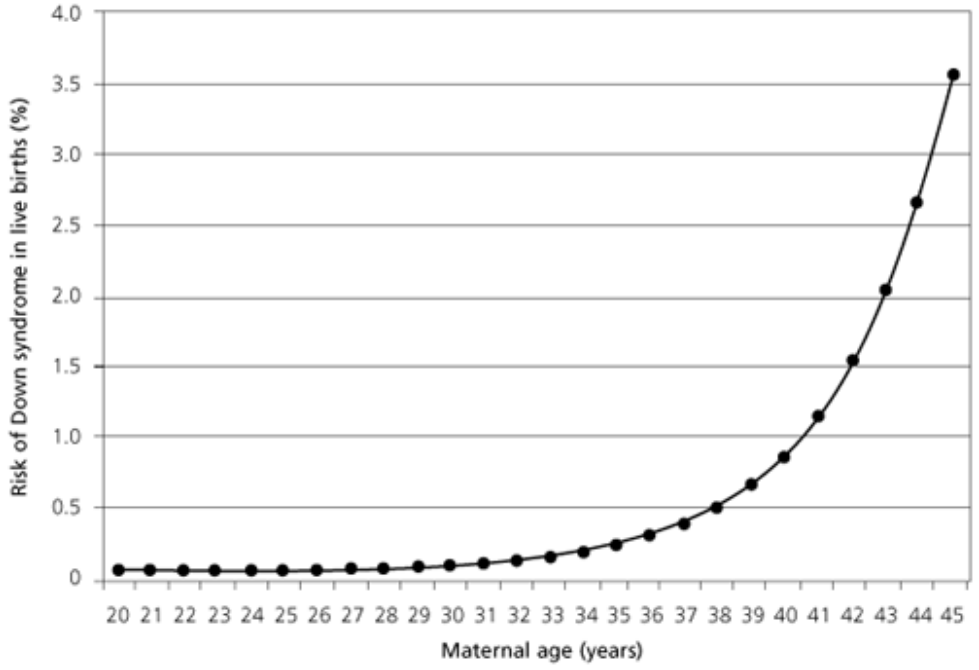


Figure 1.1. Estimated risk of Down syndrome according to maternal age.

Disappointingly, 20 years of screening in the UK using maternal age alone failed to produce a noticeable effect on the birth incidence of Down syndrome [8].

1.2.2. Recurrence Risk and Family History

If a patient has had a Trisomy 21 pregnancy in the past, the risk of recurrence in a subsequent pregnancy increases to approximately 1% above the baseline risk determined by maternal age. Diagnosis of a chromosome-21 translocation in the fetus or newborn is an indication for karyotype analysis of both parents. If both parents have normal karyotypes, the recurrence risk is 2% to 3% [11].

1.2.3. Maternal Serum Screening

1.2.3.1. First Trimester Biochemical Screening. PAPP-A and free β-hCG are two serum markers used in screening for Down syndrome in the first trimester [12, 13]. PAPP-A levels are reduced in affected pregnancies while free β-hCG levels are raised. Adding maternal age to PAPP-A and free β-hCG gives a DR of 60% and a FPR 5%, using a risk cut-off level of 1 in 250 (i.e. any woman with a risk greater than 1 in 250 is defined as high risk, and offered invasive testing).

1.2.3.2. Second Trimester Biochemical Screening. Alpha-fetoprotein (AFP), unconjugated estriol (uE3) and human chorionic gonadotropin (hCG) are the serum markers most widely used to screen for Down syndrome [14]. This combination is known as the "*triple test*". With Trisomy 21, second-trimester maternal serum levels of AFP and unconjugated estriol are about 25 percent lower than normal levels and maternal serum hCG is approximately two times higher than the normal hCG level [15].

The available second trimester screening tests are the Double, Triple and Quadruple Tests. They are compared in Table 1.1.

Table 1.1. Second trimester biochemicals.

Test	**Markers**	**DR**	**FPR**
Double	Age + AFP + hCG	59%	5%
Triple	Age + AFP + hCG + uE3	63%	5%
Quadruple	Age + AFP + hCG + uE3 + inhibin A	72%	5%

For women 35 years or older, maternal serum screening can provide an individual estimate of the likelihood of fetal Trisomy 21 [16]. However, the triple test fails to detect 10 to 15 percent of Trisomy 21 pregnancies in women in this older age group [17]. Therefore, current U.S. practice standards indicate that for women 35 years or older, maternal serum screening should not be offered as an equivalent alternative to amniocentesis or chorionic villus sampling [17–19]. Guidelines published by the American College of Obstetricians and Gynecologists state that maternal serum screening may be offered "as an option for those women who do not accept the risk of amniocentesis or chorionic villus sampling or who wish to have this additional information prior to making a decision about having amniocentesis [19].

1.2.4. Ultrasound Screening

Thirty percent of fetuses with Trisomy 21 have a major structural malformation. Congenital cardiac anomalies are the commonest (up to 40%) and of these atrioventricular canal defects and ventricular septal defects are the most frequent.

Trisomy 21 in the second trimester is also associated with nasal bone hypoplasia, increased nuchal fold thickness, duodenal atresia, echogenic bowel, mild hydronephrosis, shortening of the femur or humerus, sandal gap, and clinodactyly or midphalanx hypoplasia of the fifth finger.

1.2.4.1. First Trimester Screening. Ultrasound measurement of nuchal translucency has been studied alone and in combination with new biochemical markers as a potentially

useful first-trimester screening test for Trisomy 21. Estimates are that first-trimester screening by means of maternal age and measurement of nuchal translucency could provide a Trisomy 21 detection rate of 63 percent, with a 5 percent false-positive rate [21].

1.2.4.2. Combined Test. Recent advances include using a combination of NT and bio-chemical markers. The combination of first trimester free β-hCG, PAPP-A, NT and maternal age is known as the Combined Test, and is measured between 11 and 13 weeks. This has been reported in some studies to have a DR of 80-89% with a FPR of 5% [23, 24].

1.2.4.3. Integrated Test. The Integrated Test is the most recent screening test for Down syndrome [25]. This combines maternal age with the following:

- 11-14 weeks: NT + PAPP-A
- 15-22 weeks: AFP + hCG + uE3 + Inhibin A

The performance of this test is reported to be better than that of all others. The model of screening described by Wald and Hackshaw [23] has the major theoretical advantage of a high DR of 94% for a FPR of 5% or alternatively 85% DR with a 1.2% FPR.

Table 1.2. Milestones in the history of screening for Down syndrome.

Year	**Milestone**
1933	Association between maternal age and Down syndrome noted
1959	Trisomy 21 identified as the cause of Down syndrome
1966	First chromosome analysis from amniotic flid
1968	Prenatal diagnosis of Down syndrome
1972	Raised amniotic fluid AFP associated with open neural tube defects
1977	Maternal serum AFP screening for open neural tube defects
1988	Triple test
1991	Nuchal translucency

1.3. Literature Review

Evaluating prenatal risk of Down syndrome with non-invasive methods has been investigated over the years and still attracting academicians as an emerging research field. Existing studies heavily focus on statistical relationships between clinical variables and pregnancy outcome. These studies provide valuable information for the risk assessment of Down syndrome. However, because of the difficulty faced in manual observation of multiple variables and examination of nonlinear correlations between features, detection process requires more advanced data analysis and prediction models. On the contrary to the emergence and importance of intelligent decision support systems in Down syndrome detection process, the related literature is limited.

Brock and Sutcliffe [26] found in 1972 that the level of alpha fetoprotein in amniotic fluid increased when the fetus had a neural tube defect, and from the 1980s, the maternal serum alpha fetoprotein test began to be used for screening fetal anomalies in pregnant women.

In 1984, Merkatz et al. [27] found that the risk of Down syndrome was high when the level of serum alpha fetoprotein was low during the second trimester, but this finding alone was not sufficient for using alpha fetoprotein as an accurate Down syndrome marker [27].

Later, double marker test that added human chorionic gonadotropin (hCG) test, and triple test that added also estriol (E3) to the double test were introduced as Down syndrome screening tests in the second trimester [28]. Recently, quad test that added inhibin A was developed [28] for higher accuracy of screening.

There were large-scale prospective studies that compared accuracy among a number of multiple markers in the U.K. (Serum, Urine and Ultrasound Screening Study; SURUSS) [29] and the U.S. (First and Second Trimester Evaluation of Risk for Fetal Anueploidy; FASTER) [30], and in both studies the detection rate of quad markers was reported to be 81%.

The thickness of nuchal translucency in the first trimester is related to fetal haploidy, in particular, to Down syndrome, apart from maternal serum markers [31]. Accordingly,

the accuracy of screening can be enhanced through the first trimester combined test that measures the thickness of nuchal translucency in addition to serologic tests that measure PAPP-A and free β-hCG [31].

Wald et al. [32] proposed integrated test, which uses information on first and second trimester markers in sequence. They expected that if nuchal translucency and serum PAPP-A are measured in 10-13 weeks, and alpha fetoprotein, total hCG, estriol and inhibin A in 15-18 weeks, Down syndrome can be detected at a rate of 94% with a false positive rate of 5%, or 85% with a false positive rate of 1% [33].

In the results of SURUSS [34] and FASTER [35], the integrated test was most accurate as a Down syndrome screening test. However, the integrated test has a number of shortcomings to be an alternative general screening test in prenatal examination. First, most of pregnant women who receive a screening test in the first trimester want the termination of pregnancy immediately if abnormalities are found in the fetus, and it is safer to terminate pregnancy in the first trimester. Second, it is hard to distinguish pregnant women who cannot be followed up after first trimester serologic tests. In such a case, there could be the legal risk of not telling the results of the first trimester test, so it was not an adequate alternative at present.

More recent studies have examined the role of first trimester ultrasound markers other than NT. They suggest that absence of the nasal bone, increased impedance to flow in the ductus venosus (DV) and tricuspid regurgitation (TR) are highly sensitive and specific first trimester markers for Trisomy 21 [37-39].

In 2005, Nicolaides et al [40] proposed a two-stage screening process in the first trimester. They suggested using the combined test to triage women into high risk (1 in 100 or greater), intermediate risk (between 1 in 101 and 1 in 1000) and low risk (less than 1 in 1000). Intermediate risk women were offered further assessment of risk by first trimester ultrasound examination to determine the presence or absence of the nasal bone, presence or absence of TR and normal/abnormal doppler velocity waveform in the DV. They concluded that using this approach, more than 90% of Trisomy 21 fetuses can potentially be identified in the first trimester, for a FPR rate of 2-3%.

Table 1.3. Direct comparative data for the first and second trimester Down syndrome screens from the prospective FASTER and SURUSS trials [36].

		FASTER [35]					**SURUSS[34]**				
		FPR(%) for DR of			DR(%) for FPR of		FPR(%) for DR of			DR(%) for FPR of	
		75%	85%	95%	1%	5%	75%	85%	95%	1%	5%
1st. trimester	NT only	8.1	20	55	54	68	12.9	25	55	33	60
	PAPP-A + f-βhCG	7.1	16	42	46	67	5.5	12	33	52	74
	NT + PAPP-A + f-βhCG	1.2	3.8	18	72	85	2.3	6.1	22	66	83
2nd. trimester	Triple (AFP+hCG+E3)	7	14	32	45	69	2.9	7.1	22	51	74
	Quad (Triple+inhibin A)	3.1	7.3	22	60	81	2.6	6.1	18	63	83
1st. +2nd.	PAPP-A + Quad	1.2	3.6	15	70	86	0.8	2.7	12.5	77	90
	PAPP-A + Quad + NT	0.2	0.6	4.0	87	95	0.3	1.2	7.2	84	95

The continuing debate as to whether screening should be performed solely in the first trimester or should incorporate second trimester markers remains mostly unanswered with no prospective randomized trials to compare first versus second trimester screening. One of the main focus for screening is to achieve a high DR with a low FPR rate. It would appear that the integrated test may be the most effective test available at present. However, new first trimester markers, such as fetal nasal bone hypoplasia and TR, are being evaluated. These may prove even more effective.

Since there are no public Down syndrome databases, all of the studies mentioned above perform experiments on different proprietary datasets. A direct comparison of reported results is not possible due to the varieties of outcome measure, data features, dataset sizes, methodologies and performance criterias.

1.4. Thesis Outline

This dissertation is organized as follows:

Chapter 1 is the introductory part presenting research overview, explanation of the entire Down syndrome risk assessment process together with the widely used characteristics and methods and a literature review on Down syndrome.

We present the problem statement and relevant research questions in Chapter 2 considering the described challenges and explained relevant background information.

We propose solutions for each research question in Chapter 3 to be a based for our methodology and experiments.

Chapter 4 presents the brief definitions of the machine learning algorithms as the methodology of our study. The experiments and results are given in Chapter 5. We show the probabilistic classifiers performance in this section.

Finally in Chapter 6, we provide an overall conclusion and discussion of the future research directions.

2. PROBLEM STATEMENT AND RESEARCH QUESTIONS

In this chapter we discuss our research questions with related problem statement and background. We mainly state four research questions with additional considerations.

2.1. Characteristics of Down Syndrome Data

We need to analyze the predictor factors characterizing the outcome of Trisomies 21 in order to provide a reliable prediction model.

2.1.1. Predictive Factors

Antenatal screening and patient related data have been widely investigated as predictor factors characterizing the Down syndrome prediction as discussed in Section 1.3. The studies reporting lower prediction performance either question the sufficiency of information content of their datasets or point out investigation of new predictor features as future work since improving the information content of datasets provides better recognition performance in machine learning applications.

2.1.2. Mixed Data Type

The prognostic factors in Down syndrome dataset include both continuous (e.g. age) and categorical variables (e.g. race). Transformation of categorical variables into numeric values or discritization of continuous variables is crucial for the specific classification algorithms. Defining the most proper method for transformation produce better prediction results. The mixed data type characteristics of the datasets have been another important challenge in our research.

2.1.3. On the Effect of Pre-Processing or Post-Processing

Each real world application of standard machine learning algorithms require careful pre-processing of input data, necessary modifications to learning algorithms and post-processing of the results if necessary.

2.1.4. Imbalanced Class Distribution

Our datasets represents an imbalanced distribution of class samples. Positive Trisomies 21 outcome distributions are 0.7% and 0.9% for our datasets. This is a major effect to reduce classification performance. Besides this one of our datasets has very few instances, 213. Neither the small size nor the imbalanced class distribution problems are major factors to overcome during training phase of our classification methods.

2.2. Research Questions

In this section we define our research questions based on the standard problems and relevant background presented in the previous section.

2.2.1. Research Question 1: How Can We Construct an Efficient Non-invasive DS Prediction Model?

Trisomy 21 prediction is a typical problem of decision making under uncertainty conditions because of the various factors affecting the outcome. We have stated the Down syndrome prediction problem based on supervised learning approach. Rather than a comparison to previous studies, our objective is to build a novel applicable decision

support system for all stages of prenatal period by using advances in machine learning methods. Predicting Trisomy 21 is the preliminary study of this research.

2.2.2. Research Question 2: How can We Enhance the Methodologies to Improve the Prediction Performance?

Our goal is to decide the best pre- and post-processing techniques to handle the imbalanced class distributions and mixed data type. We analyze the assumptions of the standard machine learning algorithms, compare the common pre- and post-processing techniques and propose modifications to improve the prediction performance in Down syndrome domain.

2.2.3. Research Question 3: Is It Possible to Evaluate the Prenatal Risk With a Probabilistic Classifier Method Accurately?

In this part of our research, we apply probabilistic classification methods over the Down syndrome datasets. We use classification algorithms based on Bayes' theorem, namely Bayesian Networks and Naive Bayes.

2.2.4. Research Question 4: Does Proposed Method in Determining DS Outperform Existing Methods?

We aim to show the prediction performance of probabilistic classifiers versus other widely used classifiers in a straightforward way in this section. To represent the results clearly of the performed experiments, we use ROC analysis and performance comparison tables including accuracy and false positive rate as main criterias.

3. PROPOSED SOLUTION

We outline our proposed solutions for each research question to be a base for the methodology and experiments.

3.1. Down Syndrome Prediction as a Supervised Binary Classification Problem: Research Question 1

A learning based predictor model that makes use of artificial intelligence notion can automatically analyze large medical databases to train predictor models and provide future implications. Specifically for the Trisomy 21 prediction problem, these models can be used to predict the Down syndrome case when relevant prognostic features are supplied as model inputs.

Quality of an intelligent learning based system depends on three main factors. First, construction of a comprehensive dataset that represents the underlying characteristics of the application domain enables accurately learning the relations between input and output. Second factor is selection of best fitting models for the specific domain together with unbiased training and testing strategies that avoid the sampling and learning bias. Third factor is careful application of the model specific pre-processing techniques and necessary algorithmic modifications to enhance the prediction performance.

Initially, we have obtained two different datasets from previous researches with similar variables and characteristics. Each patient is represented as a data feature vector including 31 and 24 clinical variables and a class label: 1 for Positive Down syndrome case and 0 for negative Down syndrome case.

3.2. Handling the Imbalanced Class and Mixed Data Type Problem in Datasets: Research Question 2

The results of the initial experiments on Down syndrome prediction motivated us to improve the performance of classification. There are two ways to improve the performance of a classification task: to improve the algorithm to better fit the problem or to improve the information content of the data. Regarding our second research question, we performed experiments to improve the performance by handling the imbalance and mixed data type problems.

Learning from imbalanced datasets has been an important research interest in the last decade [40,41]. Various sampling strategies have been proposed to deal with the problem of imbalance class distribution [42-44]. On the other hand, recent studies show that adjusting the decision threshold of classifiers produce similar results with artificially changing the distribution of the instances in the training set [45,46]. We apply under- and over-sampling strategies to re-balance the dataset and adjust the decision threshold to improve the classification results.

Analysis and pre-processing of mixed datasets including a combination of continuous and categorical variables is investigated widely [47-50]. In this research, we analyze the performance of Down syndrome prediction on mixed datasets using SVM method. In addition, we use different encoding techniques for transformation of categorical variables.

3.3. Bayesian Networks and Naive Bayes for Classification of Down Syndrome: Research Question 3

In this research, we apply Bayesian Networks and Naive Bayes for modelling Down syndrome prediction as probabilistic classifiers. Bayesian Networks classifier has been popular tools for medical decision support systems in the last decade [51,52]. Specific applications include bypass surgery survival prediction [53], ovarian cancer diagnosis [54],

diagnosis of female urinary incontinence [55], diagnosis and treatment of ventilator-associated pneumonia [56] etc.

The visualization of statistical cause-effect relationships in a network structure makes the Bayesian Networks easy to understand and apply in medical applications. The components of the Bayesian Networks, i.e. nodes, arcs and conditional probabilities correspond to prognostic variables, dependencies and statistical inference, respectively. Such a model is useful especially when we need to know the underlying reason for the prediction outcome rather than a black-box model in which the explanation for the prediction is difficult to understand.

As a second method, we apply Naive Bayes for Down syndrome prediction. Naive Bayes is a simple probabilistic classifier based on Bayes' Theorem with strong independence assumptions. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can work with the Naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

An advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. When we consider our second dataset which has only 213 instances, Naive Bayes' this opportunity has a crucial role for our study.

3.4. Existing Classification Methods vs. Bayesian Methods: Research Question 4

After developing a probabilistic classification method, we mainly want to show the performance of probabilistic classifiers with respect to the non-probabilistic ones. To accomplish this we also implement well known and widely used classification methods within our research. To evaluate the probabilistic classifiers' performance clearly this step

has a major role. This comparison is based on performance metrics such as accuracy, Receiver Operating Characteristics (ROC), False Positive Rate (FPR), etc.

To evaluate probabilistic classifiers performance, we apply SVM, MLP, DT, kNN classification methods. Although, there are more classification algorithms, we implemented these ones as the most famous ones within our study. Actually in our research we mainly concentrate on how probabilistic classifiers perform well. With this comparison we answer this question objectively.

4. METHODOLOGY

In this chapter, we provide the theoretical background of the statistical and machine learning techniques that we used in our experiments. We also discuss the relevance of the selected methods to our research questions regarding the characteristics of the Down syndrome domain.

4.1. Experimental Design

The clinical studies can be categorized as retrospective and prospective according to data collection method and occurrence of events of interest. The definitions of the terms ‘retrospective’ and ‘prospective’ are given [53] as:

Retrospective: “All events of interest have already occurred and data are generated from historical records and from recall.”

Prospective: “Data collection and the events of interest occur after individuals are enrolled (e.g. clinical trials and cohort studies).”

In our research, we mainly use prospective data consisting after an enrolment process. We obtained two datasets from two different previous researches. We know the exact lifecycle of the data collection. Since the details of the data collection is not in our scope, we do not mention about it. All patients informed consent was obtained.

Prospective studies provide more robust, consistent and reliable results avoiding the potential biases in the historical data. Prospective validation of a prediction model in medical domain is necessary. Therefore, our research input data consist of prospective data, and this validates our experiments and results.

4.2. Datasets

In this research we used two different datasets from geographically two different locations. Actually there is no similarity between datasets in terms of location. By the help of this property we can validate our own experimental results and have a different point of view.

4.2.1. Dataset 1: George Washington University, United States

The research [55] of first dataset was conducted as a multicenter study of screening for Trisomies 21 and 18 among patients with pregnancies between 74 and 97 days of gestation, based on maternal age, maternal levels of free beta human chorionic gonadotropin and pregnancy-associated plasma protein A, and ultrasonographic measurement of fetal nuchal translucency. Screening was completed in 8216 patients with singleton pregnancies.

The study was approved by the institutional review board at each of the 12 participating prenatal diagnostic centers. All participants gave written informed consent. Major exclusion criteria included multiple gestation, recent vaginal bleeding equivalent to a menstrual period, pregestational diabetes mellitus, and pregnancy resulting from a donor oocyte. Patients with indications for prenatal diagnosis other than a risk of Trisomy were also excluded.

Table 4.1, Table 4.2, Table 4.3 and Table 4.4 summarize the general characteristics of George Washington University dataset. In this dataset there are 61 Down syndrome cases. There are more Down syndrome cases with respect to second dataset while the percentage of the positive Down syndrome outcome is less than the second dataset.

Table 4.1. George Washington University dataset summary.

Name	Value
Instances	8216
Features	31
Nominal Features	3
Numeric Features	28

Table 4.2. Characteristics of 8216 pregnant patients who underwent first-trimester screening.

Characteristics	Value	%
Maternal age		
16-24 years	268	3
25-29 years	1013	12
30-34 years	2815	34
35-39 years	3280	40
≥40 years	840	10
Mean maternal age	34.5	
Maternal race or ethnic group		
Black	352	4
White	6815	83
Hispanic	452	6
Asian	428	5
Other	169	2
Gestational age at screening		
74-76 days	478	6
77-83 days	2571	31
84-90 days	3223	39
91-97 days	1935	24
98 days	9	<1
Mean gestational age	85.7	

Table 4.3. BUN dataset contents.

Number	Attribute Label	Attribute Code
1	CROWN RUMP LENGTH	A1
2	GESTATIONAL AGE AT U/S (DAYS)	A2
3	FREE BETA HCG	A3
4	PAPP-A LEVEL	A4
5	MAT. AGE AT BUN SCREEN (YRS)	A5
6	AGE SPECIFIC RISK	A6
7	RACE	A7
8	MATERNAL WEIGHT (LB)	A8
9	MAT. AGE AT TERM (YRS)	A9
10	PATIENT CURRENTLY SMOKES?	A10
11	NUCHAL TRANSLUCENCY (MM)	A11
12	GESTATIONAL AGE AT BLOOD DRAW (DAYS)	A12
13	HCG MOM – UNADJUSTED	A13
14	PAPP-A MOM – UNADJUSTED	A14
15	1ST TRI. FREE BETA HCG (MOM)	A15
16	1ST TRI. PAPP-A (MOM)	A16
17	GA FACTOR IN RISK CALC (T21)	A17
18	RISK AT TERM	A18
19	1ST TRI RISK T-21	A19
20	LIKELIHOOD RATIO FOR COMBINED T21 RISK	A20
21	T-21 RISK-BIOCHEM ONLY	A21
22	T-21 RISK-NT ONLY	A22
23	1ST TRI RISK	A23
24	GA (WKS) AT RISK ASSESSMENT	A24
25	1ST TRI. NT (MOM)	A25
26	LOG10 NT MOM	A26
27	LOG10 HCG MOM	A27
28	LOG10 PAPPA MOM	A28
29	LOG10 UNADJUSTED HCG MOM	A29
30	LOG10 UNADJUSTED PAPPA MOM	A30
31	DS	A31

Table 4.4. BUN data numeric variables statistics.

	Max	Min	Mean	StdDev	Missing	Distinct	Unique
MAT. AGE	46	16	33.484	4.58	0 (0%)	31	0 (0%)
CRL	86	38	59.358	11.315	0 (0%)	458	16 (0%)
GA_US	98	74	85.743	5.69	0 (0%)	25	0 (0%)
HCG	491	2.92	59.425	32.848	0 (0%)	7287	6466 (79%)
PAPP-A	616	0.1	3.619	7.977	0 (0%)	7610	7049 (86%)
WGTLB	357	82	145.678	29.686	2 (0%)	195	21 (0%)
AGE_EDC	47.4	16.7	34.508	4.578	0 (0%)	4499	2316 (28%)
NUCHAL	10.6	0.1	1.527	0.6	0 (0%)	65	16 (0%)
GA_DRAW	97	64	84.627	6.176	0 (0%)	34	0 (0%)
UNAJHCGM	9.66	0.05	1.122	0.606	0 (0%)	351	56 (1%)
UNAJPAPM	210	0.06	1.127	2.809	0 (0%)	375	115 (1%)
HCGMOM	8.46	0.06	1.191	0.623	0 (0%)	365	60 (1%)
PAPMOM	207	0.06	1.216	2.762	0 (0%)	388	117 (1%)
FACTOR21	1.8	1.34	1.422	0.044	0 (0%)	55	3 (0%)
TERMRISK	1567	13.1	517.412	379.468	0 (0%)	4351	2443 (30%)
FTRSK21	1151	9	364.226	267.196	0 (0%)	1069	89 (1%)
LRC_T21	764	0.01	0.866	9.803	0 (0%)	2471	1998 (24%)
BIORSK21	8115	5	1566.07	1725.53	0 (0%)	2645	1078 (13%)
NTRSK21	8115	5	1829.331	1691.76	0 (0%)	3122	1329 (16%)
RCALCT21	10000	5	2788.888	2571.24	0 (0%)	2670	1208 (15%)
GA_CALC	19	10	13.017	1.031	1314 (16%)	10	2 (0%)
NTMOM	5.77	0.09	0.988	0.372	0 (0%)	263	70 (1%)
LOGNTMOM	0.76	-1	-0.029	0.141	0 (0%)	263	70 (1%)
L_HCGMOM	0.93	-1.2	0.015	0.243	0 (0%)	365	60 (1%)
L_PAPMOM	2.32	-1.2	0.005	0.222	0 (0%)	388	117 (1%)
L_UNHCGM	0.99	-1.3	-0.015	0.252	0 (0%)	351	56 (1%)
L_UNPAPM	2.32	-1.2	-0.036	0.237	0 (0%)	375	115 (1%)

4.2.2. Dataset 2: Trakya University, Turkey

The research [54] of this dataset was performed in Trakya University Faculty of Medicine, Department of Obstetrics&Gynecology, on 213 consecutive pregnant women aged between 18 and 43 years admitted for antenatal care at 11-14 weeks of gestation. Twins or higher order pregnancies, pregnancies ending in sponaneous abortion or with congenital anomalies detected at the first trimester and patients that did not deliver in the clinic or were lost during follow-up were excluded from that study. All patients were

delivered in the department and the newborns were examined after birth for possible anomalies. That study was approved by the Ethics Committee for Human research at Trakya University, Turkey. And informed consent was obtained from all patients. The study population consisted of Turkish women living in the Trakya Region of Turkey.

During data collection process, age, maternal smoking habit, previous fetuses with anomalies, presence of diabetes were noted, height and weight were obtained and body mass index calculated from all women. A detailed structural survey by ultrasound was performed on each fetus. Crown rump length (CRL), NT and DV flow patterns were measured by the same clinician during periods without uterine contractions and in the absence of fetal body movements. Three measurements for NT were obtained and the highest was accepted for calculation of risk for the combined test. Blood samples were obtained from the subjects through venipuncture to perform the PAPP-A and f β-HCG assays. All values were calculated by multiples of median (MoM) according to gestational age.

In screening program, marker levels are described in terms of Multiple of the Median (MoM). This is to allow for the fact that marker levels vary with gestational age. MoM values are calculated by dividing an individual's marker level by the median level of that marker for the entire population at that gestational age in that laboratory. Using MoM values, rather than absolute levels, also allows results from different laboratories to be interpreted in a consistent way.

Table 4.5, Table 4.6 and Table 4.7 summarize the general characteristics of Trakya University dataset. In this dataset there are only 2 positive Down syndrome outcome. This issue of fact make the prediction difficult. With 2 positive samples classification methods could not be trained very well. To negotiate this problem we use some techniques.

Table 4.5. Trakya University dataset summary.

Name	**Value**
Instances	213
Features	24
NominalFeatures	4
Numeric	20

Table 4.6. Trakya University dataset content.

Number	Attribute Label	Attribute Code
1	Age	B1
2	Age Specific Risk	B2
3	BMI	B3
4	Smoke	B4
5	DateLstMenstrual_PW	B5
6	USG_PW	B6
7	CRL	B7
8	NT	B8
9	PI	B9
10	FH_R1	B10
11	FH_R2	B11
12	MthdBirth	B12
13	BirthWeight	B13
14	Gender	B14
15	PAPP	B15
16	BhCG	B16
17	PAPP_MoM	B17
18	BhCG_MoM	B18
19	Risk2	B19
20	AFP_MoM	B20
21	hCG_MoM	B21
22	uE3_MoM	B22
23	Risk3	B23
24	DS	B24

Table 4.7. Trakya University dataset numeric variables statistics.

	Max	Min	Mean	StdDev	Missing	Distinct	Unique
Age	18	43	27.864	4.961	0 (0%)	24	3 (1%)
BMI	17.2	39.5	23.623	4.079	0 (0%)	104	56 (26%)
DateLstMenstrual	11	14	12.439	0.717	0 (0%)	23	2 (1%)
USG_PW	11	14	12.389	0.703	0 (0%)	22	2 (1%)
CRL	40	81	58.562	9.176	0 (0%)	150	112 (53%)
NT	0.59	3.6	1.17	0.306	0 (0%)	65	19 (9%)
PI	0.62	1.31	1.055	0.134	1 (0%)	51	8 (4%)
FH_R1	145	187	162.437	8.004	0 (0%)	22	5 (2%)
FH_R2	145	192	164.371	7.401	0 (0%)	27	8 (4%)
BirthWeight	1090	4260	3271.469	444.094	2 (1%)	100	48 (23%)
PAPP	0.64	9.9	2.769	1.637	0 (0%)	65	25 (12%)
BhCG	3.5	238	42.106	35.516	0 (0%)	182	154 (72%)
PAPP_MoM	0.18	2.4	0.813	0.398	0 (0%)	104	48 (23%)
BhCG_MoM	0.11	8.74	1.632	1.33	0 (0%)	142	96 (45%)
AFP_MoM	0.16	2.77	1.043	0.368	28 (13%)	101	54 (25%)
hCG_MoM	0.23	3.02	1.157	0.58	28 (13%)	115	67 (31%)
uE3_MoM	0.16	4.04	1.196	0.456	28 (13%)	105	55 (26%)

4.3. Dimension Reduction

Data mining applications deal with huge amount of data. The time and space complexity of any classifier or regressor directly depends on the input data size [56]. Dimensionality reduction techniques can be applied to the input data to obtain a reduced representation of the original dataset without losing the integrity of the original data [57]. However, all features may not be necessarily relevant to the outcome. In some cases, a reduced feature subset would better represent the information content of the underlying dataset and overcome “curse of dimensionality” in learning phase of the classification algorithm. Dimension reduction can be generally divided into two techniques [58]: feature selection is the technique of selecting a subset of relevant features for building robust learning models and feature extraction transforms the input data into a new set of features that are the combinations of the original variables.

Reducing the dimension of input data by eliminating the redundant information helps to:

- Improve the performance of prediction
- Reduce the computational complexity of the classification algorithms
- Prevent storage of unnecessary medical data
- Provide better understanding of the underlying process
- Simplify the utilization of the model in the clinical routine

As a pre-processing step, in our study, we try to obtain a strong input with smaller volume without losing accuracy. Thus, reducing the dimension and revealing the underlying information can be quiet important in this work. We apply both techniques, feature selection and feature extraction by different type of algorithms.

4.3.1. Feature Selection

Feature selection strategies are applied to explore the effect of irrelevant features on the performance of classifier systems [59]. In this phase, an optimal subset of features which are necessary and sufficient for solving a problem is necessary. From a theoretical perspective, it can be shown that optimal feature selection for supervised learning problems requires an exhaustive search of all possible subsets of features of the chosen cardinality. If large number of features is available, this is impractical. For practical supervised learning algorithms, the search is for a satisfactory set of features instead of an optimal set.

In our research, we use two datasets. As shown in the Table 4.1 and 4.5 we have 31 and 24 features respectively, which means there are 2^{31} and 2^{24} possible subsets for feature selection techniques. Testing all the subsets is not feasible computationally. In order to reduce the search space we need to apply some heuristics such as Information

Gain feature weighting approach. We aim to select the most relevant k features of d dimensional features and discard the unnecessary (d-k) ones.

Many approaches have been proposed for feature selection such as well known methods decision tree as filter and recursive feature elimination with support vector machine as wrapper. SVM-RFE aims to minimize cost function as performance measure [60], on the other hand DT utilizes tree induction algorithm with the entropy as an evaluation measure [61].

4.3.1.1. Recursive Feature Elimination with Support Vector Machine. Recursive Feature Elimination (RFE) is a wrapper method that utilizes the generalization capability embedded in SVM. RFE keeps the independent features containing the original dataset information while eliminating weak and redundant features [62]. However, the subset produced by SVM-RFE is not necessarily the ones that are individually most relevant. Only taken together the features of a produced subset are optimal informative [63].

The working methodology of SVM-RFE is based on backward selection where algorithm starts with whole features and iteratively eliminates the worst one until the predefined size of the final subset is reached. At each iteration, the remaining features must be ranked again [64].

SVM-RFE working principles at each iteration could be examined in three steps:

(i) Training the classifier (SVM)
(ii) Computing the ranking criterion for all features
(iii) Removing the feature with smallest ranking criterion

```
Function RFE-SVM(TD, AF, RS)
Initialize
        TD: Training data
        AF: All features in the dataset
        RS: Reduced feature subset
Begin
        While(number of AF > RS)
                Train SVM on TD with the feature space AF
                Rank the features of F in the descending order
                RFS:= AF – {feature with the smallest rank in AF}
                AF = RFS
```

Figure 4.1. FRE-SVM algorithm.

For ranking criterion, there are different algorithms in the literature such as entropy [65] or square of the weight of separating hyperplane (w^2) [64]. In our work, we use Weka SVM-RFE tool [66] which uses square of weight as ranking criteria where in each iteration the feature which causes minimum variation in the SVM cost function is removed from feature space. We assume that in each step, trained SVM produces weight vector w^* according to the formula below where α_i are Lagrange multipliers which are greater than zero for support vectors:

$$w^* = \sum_{i \in SV} y_i \alpha_i^* x_i \tag{4.1}$$

For the trained SVM with the weight vector w^*, the cost function is $J(w)$:

$$J(w) = \frac{1}{2}\|w\|^2 \tag{4.2}$$

In order to find the variation in cost function of SVM $\delta J(i)$:

$$\delta J(i) = \frac{1}{2}\frac{\partial^2 J(w)}{\partial w_i^2}(\delta w_i)^2 = \frac{1}{2}(w_i)^2 \tag{4.3}$$

Feature, which causes minimum variation is ranked and removed from feature space. SVM-RFE algorithm is given in Figure 4.1. In SVM-RFE, computational cost is higher while only one feature is removed in each step. When several features are removed at a time, feature subset ranking must replace with feature ranking.

4.3.1.2. Decision Tree. Decision tree is a widely used predictive model for supervised learning. DT is used for both classification and prediction. DT learning algorithm is greedy and based on "divide and conquer" approach.

Decision tree is a classifier in the form of a tree structure as shown in the Figure 4.2, where each node is either:

- a leaf node: indicates the value of the target class of examples, or
- a decision node: specifies some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test.

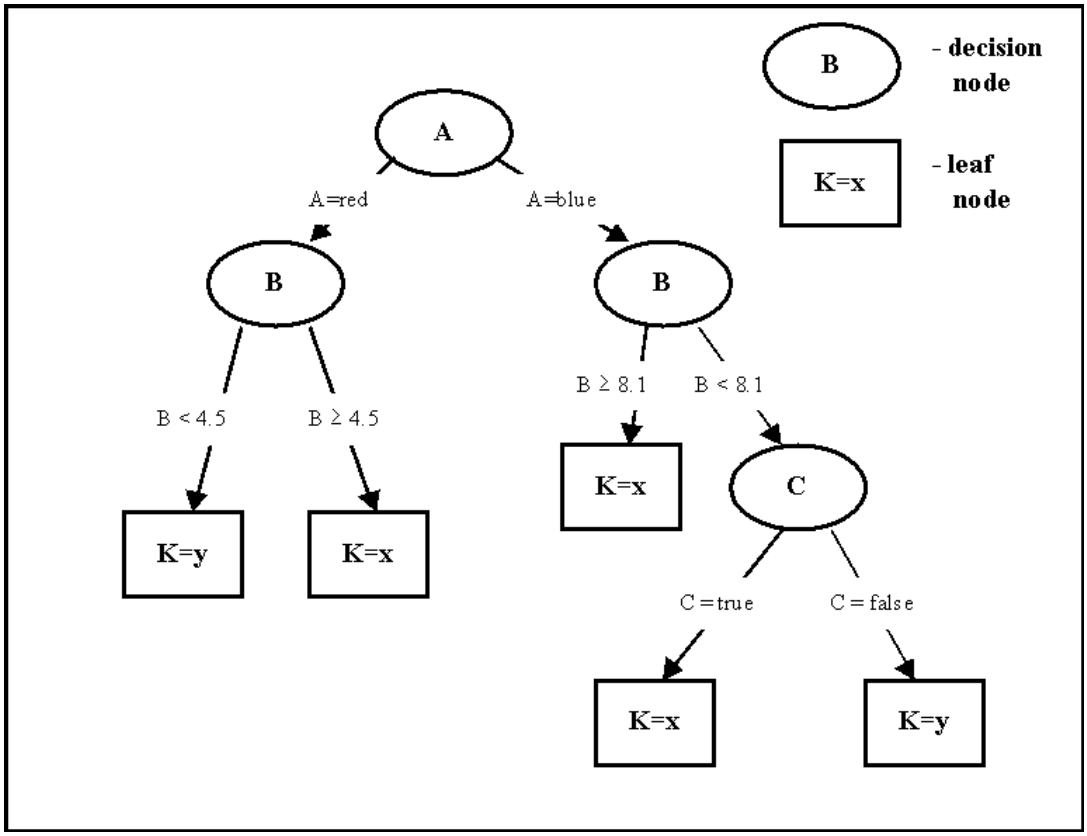


Figure 4.2. An example of a simple decision tree.

In a DT learning starts at the root node with all features by splitting the input space into two subspaces and continues recursively with the corresponding subset until a leaf node is obtained. Learning ends when the best split is reached. The measure of the good split is impurity which is determined as if all instances of the branch are labelled as the same class.

$$\widehat{P}(c_i|x,m) = p_m^i = \frac{N_m^{\ddot{u}}}{N_m} \tag{4.4}$$

The measure of impurity is *entropy* [56]. The best split is obtained when entropy is minimized. Entropy formula for node m is given in Equation 4.5.

$$I_m = -\sum_{i=1}^{k} p_m^i \log_2 p_m^i \tag{4.5}$$

Decision tree is also used as a feature selection algorithm. The final tree consists of the most relevant features and discards irrelevant ones. In our work, we use J48, which is a C4.5 tree [67] implemented in WEKA as a feature selection method. J48 is a classification tree and recursively searches the input data until it maximizes the classification performance and extracts the features that create the best splits.

4.3.1.3. Forward Feature Selection. Forward feature selection (FFS) is a subset selection method that starts with a null feature-subset and each step, it adds one feature that decreases the error most. It continues until any further addition does not decrease the error. Pseudo algorithm of the forward selection is shown in Figure 4.3.

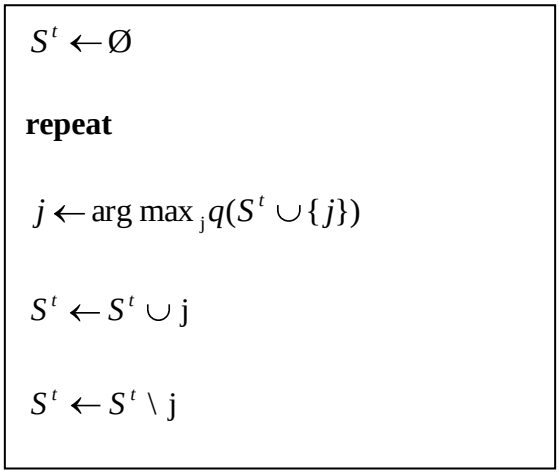


Figure 4.3. Pseudocode of forward feature selection.

4.3.2. Feature Extraction

Feature extraction aims to replace original variables by a smaller set of underlying variables. It uses linear transformation while transforming all variables to a reduced

dimension space without loss of information [68]. In recent researches, kernel and nonlinear transformation are proposed for feature extraction techniques [69-71]. Principal Component Analysis (PCA) is a widely used feature extraction algorithm that uses linear transformation. In our work, we use PCA to reduce the dimension of input data.

4.3.2.1. Principal Component Analysis. PCA transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The principal components are linear combinations of original features. Each attribute is multiplied by a coefficient, where these coefficients correspond to the elements of the principal eigenvectors.

The mathematical technique used in PCA is called eigen analysis. A solution for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products is carried out. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component.

4.4. Classification

In this section, we discuss different methods for classification that we apply within our research. Mainly we group the classification methods into two categories, probabilistic classifiers and non-probabilistic ones. We have used Naive Bayes classifier (NB) and Bayesian Networks as probabilistic classifiers. On the other hand, k-Nearest Neighbor (kNN), Decision Tree (DT), Support Vector Machines (SVM) and Multilayer Perceptron (MLP) classifiers implemented as non-probabilistic classifiers. We have chosen these classifiers because we believe that the most popular representatives of diverse algorithms (statistical classifiers, decision tree approaches, neural networks, support vector machines and nearest neighbor methods) are included [75,76]. We do not repeat the formulations of the selected classifiers in detailed here since they are well-known methods to machine learning community. We present a brief definition for these classifiers. We have performed comparison experiments of these classifiers. We aim to evaluate the performance

comparison experimentally to show the difference between probabilistic classifiers and non-probabilistic ones.

4.4.1. Naive Bayes

Naive Bayes is a simple probabilistic classifier based on Bayes' theorem, where features are assumed to be independent given the class. The assumption of independence makes it much easier to estimate these probabilities since each attribute can be treated separately. For example, an animal may be considered to be a dog if it is barking and has four legs. Even if these features depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to be the probability that this animal is a dog.

Naive Bayes algorithm works as follows: for each decision class it computes the conditional probability that decision class is the correct one, given an object's information vector. The algorithm assumes that the object's attributes are independent. The probabilities involved in producing the final estimate are computed as frequency counts from a "master" decision table [77].

Given the above description of NB, we can say that the probability of getting the string of feature values $P(X_j^1 = a_1, X_j^2 = a_2, ..., X_j^n = a_n \mid \mathrm{C_i})$ is just equal to the product of multiplying together all of the individual probabilities which is much easier to compute as well as reducing the curse of dimensionality: $P(X_j^1 = a_1 \mid \mathrm{C_i}) \times P(X_j^2 = a_2 \mid \mathrm{C_i}), ..., P(X_j^n = a_n \mid \mathrm{C_i}) = \prod_k P(X_j^k = a_k \mid \mathrm{C_i})$.

NB classifier selects the class $\mathrm{C_i}$ for which the following computation is maximum [78]:

$$P(C_i \mid \mathrm{x}) \alpha P(C_i) \prod_k P(X_j^k = a_k \mid \mathrm{C_i}) \tag{4.6}$$

Despite its simplicity, NB is successful in many applications [79]. Its advantage is that it requires a small amount of training data to estimate the parameters necessary for classification.

4.4.2. Bayesian Networks

A Bayesian network is a directed acyclic graphical model that encodes probabilistic relationships among variables of interest [80].

Bayesian Networks allow efficient representation of the joint probability distribution over a set of random variables. The network structure is used to characterize a probability distribution for each node depending on its parents. And posterior probabilities are computed in the form of local conditional distributions.

A Bayesian network is represented by $B = \langle G, \Theta \rangle$, where G is a directed acyclic graph. The nodes of the graph correspond to the random variables $X_1, X_2, ..., X_n$ which are the dataset features and edges represent direct dependencies between the associated variables. The graph G encodes the independence assumption where each variable X_i is independent of its nondescendants given its parents $\prod_{X_i}$ in G. The second component Θ represents the conditional probability distribution that quantifies the dependency between the nodes.

A Bayesian network defines a unique joint probability distribution over the set of random variables X_i in the network given by:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid \prod_{X_i}) \tag{4.7}$$

where $\prod_{X_i}$ denotes the set of parents of X_i in the network.

In practice, the components of the Bayesian Networks are unknown and must be inferred from the data. Learning a Bayesian network from data involves two subtasks, structure learning, which is necessary to identify the topology of the network, and parameter learning, that identifies the statistical parameters for a given network topology.

Most studies concentrate on structure learning which is a complex procedure when there are lots of input features [81-83]. Learning the parameters in conditional probability tables is recognized as a trivial task based on frequency counts of data points when the observed frequencies are optimal in a sufficiently large database [82]. Here, we review the main approaches for construction of the network structure and estimation of parameters when learning Bayesian networks from data.

Structure learning is a search for encoding appropriate dependencies between the features of a given dataset. It has been argued that Bayesian network structure learners are computationally expensive requiring an exponential number of conditional independence tests [82]. There are two main approaches to learn the network structure from data efficiently reducing the search space: constraint based methods and methods that maximize a selected score.

Simple learning algorithm (SLA) [82] and three-phase dependency analysis (TPDA) [82] are examples of constraint based methods that make use of information theory concept in order to reduce the computational complexity of the structure learning procedure. Reiz and Csato also propose a mutual information based approach where direct causal relations encoded by the BN are interpreted as the maximum of conditional mutual information between nodes [81].

The algorithms based on a scoring function attempt to find a graph that maximizes the selected score, which evaluates how well a given network matches the data. Different learning algorithms can be obtained depending on the definitions of the scoring function and on the search procedure used. Meloni *et al.* propose a variation of standard search-and-score approach that computes a square matrix containing the mutual information among all pairs of variables [83]. The matrix is binarized to find what relationships must be prevented. This approach prevents the inference of too many connections.

Furthermore, there are well known simple Bayesian Networks classifiers with highly constrained dependency structures: Naive Bayesian network assuming mutual

independence of the feature variables given the class variable and Tree Augmented Network (TAN) representing a tree-like dependency structure over the feature variables [84].

Parameter learning in Bayesian Networks is often based on Frequency Estimates (FE) which determines the conditional probabilities by computing the frequencies of instances from the data. The FE method is efficient since it counts each data point in the training set only once. The parameters estimated using FE method maximize the likelihood of the model given the data and thus FE is known as a generative learning method [85].

The relative frequencies in the CPT are obtained as follows:

$$\widehat{P}(X_i = x \mid \prod_{X_i} = \vec{u}) = \frac{count(X_i = x, \prod_{X_i} = \vec{u})}{count(\prod_{X_i} = \vec{u})} \tag{4.8}$$

The classification capability of FE method is argued because of the generative property. Grainer and Zhou proposed a gradient descent based discriminative parameter learning method, ELR, that significantly outperforms FE method with a high computational cost [86].

A Discriminative Frequency Estimate (DFE) is proposed to maximize the generalization accuracy of classification rather than likelihood [85]. The authors compared the DFE and FE methods based on Naive Bayesian network structure and showed that DFE significantly improve the performance of classification in terms of accuracy. However, it has been widely accepted that accuracy is not an appropriate performance measure especially for imbalanced datasets. On the other hand, the training time of DFE method is significantly higher than FE method.

4.4.3. Support Vector Machines

Support Vector Machines (SVM) is a discriminant-based method and used for both classification and regression. In classification, SVM tries to find the optimal separating

hyperlane which maximizes the distance between data points from different classes as shown in Figure 4.4.

The distance from the hyperplane on each side is called as margin and SVM tries to maximize the margin. As shown in the left side of the Figure 4.4, H3 doesn't separate the two classes. H1 does, with a small margin and H2 with the maximum margin.

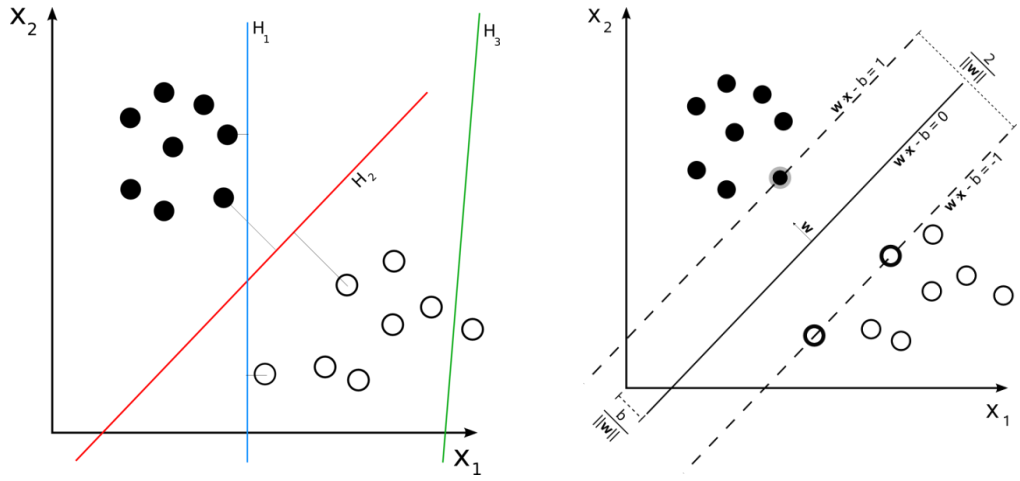


Figure 4.4. Support Vector Machine.

The key idea of SVM is to map the original input space into a higher dimensional feature space using kernel functions. Given a set of training data pairs $(x_i, y_i), y_i \in \{+1, -1\}$ final decision function is in the form:

$$f(x) = \left(\sum_i \alpha_i y_i K(x_i \cdot x) + b \right) \tag{4.9}$$

where $K(x_i \cdot x)$ is the Kernel transformation. The most popular kernel functions are:

- Linear: $K(x_i \cdot x) = x_i^T x$
- Polynomial of degree p: $K(x_i \cdot x) = (1 + x_i^T x)^p$
- Radial Basis Function: $K(x_i \cdot x) = \left[-\frac{\|x_i - x\|^2}{\sigma^2} \right]$

The optimum Kernel function and related parameters should be selected in the training phase when using SVM classification.

A penalty term C is defined as an upper bound on the Lagrange multipliers α_i trading off the complexity of the algorithm and misclassification.

$$0 \le \alpha_i \le C, \forall i \tag{4.10}$$

A higher C minimize the misclassification but may also lead over fitting of the model. Therefore the value of C needs to be tuned in the training phase in addition to Kernel parameters.

Finally, the class of an instance is decided with respect to the sign of the decision function, if $f(x) \ge 0$ then C_{+1} otherwise, C_{-1}.

SVM computes the distances of instances to the separating hyperplane in the new input space. This computation is based on assumption of continuous numerical variables. However, the dataset may include categorical features as in our Down syndrome dataset. In that case it is crucial to transform the categorical variables to continuous numerical values.

4.4.4. k-Nearest Neighbor

k-Nearest Neighbor (k-NN) algorithm is a method for classifying objects based on the closest training examples in the feature space. The measure of closeness is in terms of d

dimensional input space. There are different measurements such as Euclidean Distance or Mahalanobis Distance. Euclidean distance is a linear distance between two points which is given in Equation 4.11. Mahalanobis distance calculates the distance between two data points by the variation in each component of the points which is given in Equation 4.12.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.11}$$

$$d(x,y) = \sqrt{(x-y)\sum^{-1}(x-y)} \tag{4.12}$$

After distances between training data and new instance are calculated, k nearest neighbors are determined. Then, the class probabilities are calculated as a proportion of the number of training instances which belong to class *i* to the total number of training instances. In this work, we use Weka IBK to apply kNN on Down syndrome datasets. We prefer to consider different k numbers as closeness factor and Euclidean Distance as distance measure.

4.4.5. Multi Layer Perceptron

Multilayer Perceptron (MLP) is a nonparametric neural network structure and used for both classification and regression. Feedforward MLPs are the most widely used Artificial Neural Network (ANN) models. MLP is composed of three layers: an input layer, hidden layers and an output layer. A two hidden layer MLP is shown in Figure 4.5. In MLP, using one hidden layer is generally preferred in the case of reducing the complexity. Furthermore, large number of hidden units may cause overfitting, thus hidden layer may contain either predefined number of hidden units or optimal number of hidden units can be determined during learning.

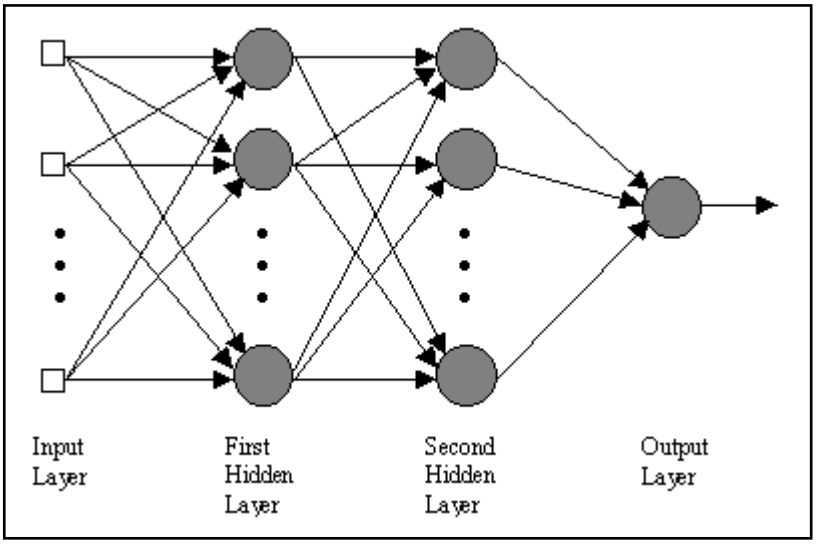


Figure 4.5. Block diagram of a two hidden layer multiplayer perceptron.

MLP learning process starts at the input layer where no calculation is applied. Briefly, hidden units nonlinearly transform the *d* dimensional input space to *h* dimensional space. The output units produce the output values as linear combinations of the *h* dimensional activation values computed by hidden units [56].

We use Weka Multilayer Perceptron function, trained by backpropagation algorithm. Back propagation algorithm updates the current values according to the predicted output value of previous layer. We use 0.8 as learning rate and maximum 500 epochs are allowed. Only one hidden layer is preferred with five hidden units. As activation function, *sigmoid* is used which is given in Equation 4.13. Sigmoid produces the output in the [0, 1] range.

$$f(u) = 1/(1+e^{-u}) \tag{4.13}$$

As output is produced by sigmoid and the problem is binary class problem, we use sigmoid(0) = 0.5 as threshold value for positive class [56]. If the produced output is greater than 0.5, patient has a positive Down syndrome case, otherwise there is no abnormalities.

4.5. Pre-Processing

In our work, we mainly use two different pre-processing operations to enhance classification algorithms and to increase experiment reliability.

4.5.1. Resampling Imbalanced Data

A common approach to overcome the problem of imbalance is to rebalance the datasets artificially. Two main resampling strategies are over-sampling that replicates instances from the minority class [43] and under-sampling where some of the instances in the majority class is removed [42].

4.5.2. Categorical Variables Conversion

Performance of distance based classifiers, such as SVM, depends on accurate transformation of categorical variables into numeric data. SVM requires each data sample to be represented as a feature vector of real numbers [88]. Therefore, categorical features should be converted into numeric values prior to classification. After transformation of categorical variables, the input data were normalized to 0 mean and standard deviation of 1.

The aim of data type transformation is to preserve the information content of the original dataset while adapting the input data to a particular analysis tool. We use binary encoding in the initial experiments of SVM classification and a frequency based encoding technique for better transformation [90].

4.5.2.1. Binary Encoding. Binary encoding maps categorical variables to higher dimensional features representing equal Euclidean distances between categories and has

been applied as a common pre-processing stage for SVM applications [88, 89].

For a particular categorical variable including N categories, each category is represented by a sequence of N bits. The i th bit corresponding to original category is set to 1 and the others are set to 0. For example, the race feature in Dataset 1 includes seven categories. When binary encoding is applied, the categories 1,2...7 correspond to 0000001, 0000010... 1000000 respectively. In this case the Euclidean distance between each category is equal, however, this may not be the actual case. Also, the input dimensionality is increased by adding dummy variables that may yield to “curse of dimensionality" in learning phase [56, 91].

4.5.2.2. Frequency Based Encoding. The literature present variances of binary encoding, frequency based and expert judgement approaches for transformation of categorical variables. However, comparative analysis of these methods is limited and also, to the best of our knowledge, there is not a generalized frequency based encoding scheme.

Johannson, et al., deal with visualization of mixed datasets and propose interactive quantization of categorical variables that incorporates information about relationships among continuous variables as well as makes use of the domain knowledge of the data analyst [92]. A Simple Correspondence Analysis (SCA) has been applied based on the frequencies of categories in the dataset.

The basic idea behind this transformation is to reflect the effect of categorical code on the outcome. The frequency of any categorical code in positive class is assumed to have positive effect while the occurrence in negative class is considered as negative effect. Hence, the new numerical value of a categorical code is defined as the difference between frequencies in positive and negative classes in the range of [-1,1].

4.6. Training and Testing Strategies

4.6.1. Dataset Splitting

In our experiments, two-thirds of the dataset was randomly selected for establishing a predictor model and the remaining one-third was utilized for testing. This random splitting was performed considering stratification principle in order to ensure that the proportions of positive and negative cases of Dowm Symdrome were the same in both training and test sets as in the original dataset. For each classifier, the model parameters were optimized on the 2/3 dataset using 10 fold cross validation strategy. The trained model was assessed on the separate 1/3 dataset to predict the class labels of the previously unseen data samples. Finally, the predictions were compared to the actual outcomes in order to evaluate the performance of the classification model.

4.6.2. k-fold Cross Validation

As stated before, our datasets are imbalanced thus we apply above classifiers with k-fold cross validation method. We have used 10 fold cross validation for parameter optimization on the training set. With 10 fold cross validation, at each iteration, input is divided into 10 partitions. Nine of them are used for training and remaining samples are used for validation. It is obvious that with 10 fold cross validation, when training is finished, each data samples is used nine times as training sample and one time as validation sample.

4.7. Performance Evaluations

In machine learning applications, the most common evaluation measure is accuracy that is the percentage of correctly predicted samples. However, in case of prediction on imbalanced datasets, accuracy is not a sufficient measure for evaluating classifiers'

performance. For example if the majority class in a dataset constitute 85% of total samples, predicting all the samples as belonging to majority class inherently yields an accuracy of 85%. Although such an accuracy level seems high, the predictor system does not provide any information about the minority class. Both the datasets we use have these type of distribution. Minority class has less than 1% proportion. Therefore, additional performance metrics are required to evaluate predictions for each class separately.

4.7.1. Performance Metrics

As best practice method, in medical machine learning applications, *sensitivity* and *specificity* measures are also widely used besides the common accuracy measure [94-96]. Formula definitions for these performance criteria are given in Equation 4.14. – 4.16. All the performance measures are derived from Figure 4.6.

Accuracy is the proportion of true results (both true positives and true negatives) in the population.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4.14}$$

On the other hand, *precision* or positive predictive value is defined as the proportion of the true positives against all the positive results (both true positives and false positives).

$$\Pr ecision = \frac{(TP)}{(TP + FP)} \tag{4.15}$$

Sensitivity of a test is the proportion of people who have the disease who test positive for it. Sensitivity relates to the test's ability to identify positive results.

$$Sensitivity = \frac{(TP)}{(TP + FN)} \tag{4.16}$$

Sensitivity = probability of a positive test given that the patient is ill

		Actual Case		
		Actual Positive	Actual Negative	
Predicted	Predicted Positive	**True Positive** **TP**	**False Positive** **FP**	**Positive predictive value** = Σ True Positive / Σ Test Outcome Positive
	Predicted Negative	**False Negative** **FN**	**True Negative** **TN**	**Negative predictive value** = Σ True Negative / Σ Test Outcome Negative
		Sensitivity = Σ True Positive / Σ Condition Positive	**Specificity** = Σ True Negative / Σ Condition Negative	

Figure 4.6. Relationships among terms.

Specificity of a test is defined as the proportion of patients who do not have the disease who will test negative for it. Specificity relates to the ability of the test to identify negative results.

$$Specificity = \frac{(TN)}{(TN + FP)} \tag{4.17}$$

Specificity = probability of a negative test given that the patient is well

4.7.2. ROC Analysis

In the machine learning community, after realization of the weakness of simple accuracy rate as a performance measure, the use of Receiver Operator Characteristics (ROC) curves [97] have gained an increasing attention. The ROC curve plots the sensitivity versus (1-specificity) by adjusting the decision threshold of classification and enables comparison of classifiers using a single performance measure that is the area under the curve (AUC) [98].

Higher sensitivity and lower false alarm (1-specificity) rates were targeted in our prediction; therefore the classifier with the largest AUC dominates the others. Figure 4.7 shows an example ROC curve where classifier 1 performs better than classifier 2 in terms of AUC.

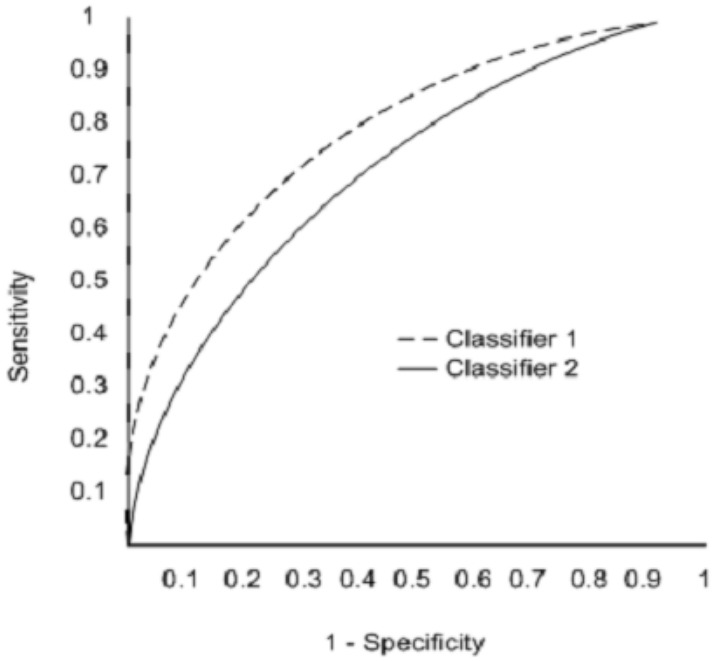


Figure 4.7. An artificial ROC curve illustrating two classifiers: Classifier 1 has larger AUC than classifier 2.

It has been shown that, the AUC represents the most informative and objective performance measure within a benchmarking context [75] especially in case of imbalanced

class distributions [45]. The datasets used in this research represents an imbalanced nature consisting of more than 99% negative and less than 1% positive cases. Hence, classifier comparison and feature subset selection have been performed according to AUC measure.

4.8. Post-Processing

Nave Bayes classifier computes the class posterior probabilities, $P(C_i \mid \mathrm{x})$ of input data (x) for both negative and positive classes. In case of binary classification, the default decision threshold was 0.5 and the patient was decided to belong to the class with the highest posterior probability.

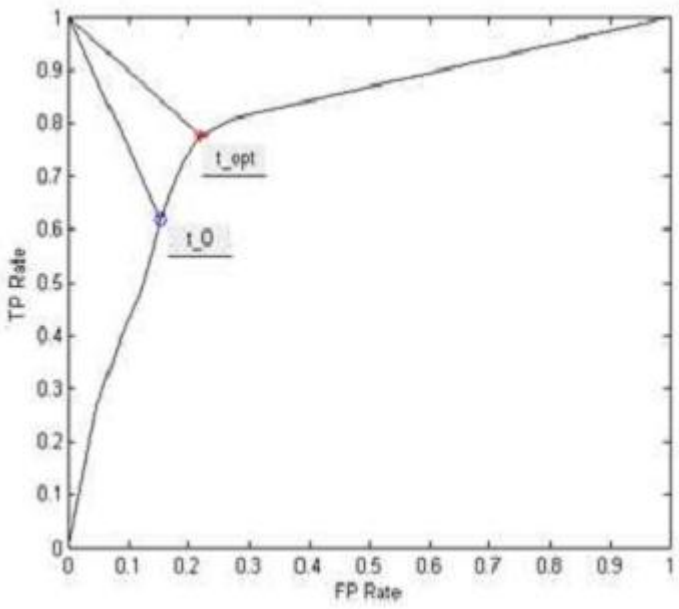


Figure 4.8. A ROC curve illustrating the effect of threshold optimization: Default threshold((t_0) and optimum threshold (t_{opt}).

The TPR and FPR have been calculated for a single threshold (default 0.5) that maps to a single point on the ROC curve. However, Provost clearly defined that, "when studying problems with imbalanced data, using the classifiers produced by standard machine

learning algorithms without adjusting the output threshold may well be a critical mistake" [46]. Since, the datasets that we have utilized in this study represent imbalanced class distributions of positive (99%) and negative (1%) classes of Down syndrome, it is necessary to evaluate the performance of classification for different thresholds. We need to determine the optimum probability threshold considering both sensitivity and false alarm rates.

It is necessary to mention critical points on the 2D ROC curve. The lower left point (0,0) represents assigning all instances to negative class. Hence, there are no positive predictions yielding TPR and FPR to be 0. Conversely, upper right corner (1,1) indicates positive prediction for all instances. The upper left point (0,1) represents perfect classification. Therefore, the threshold value that gives the nearest point to (0,1) is accepted as the optimum decision threshold (t_{opt}) in Figure 4.8. Choosing a point on the left-hand side of the (t_{opt}) reduce false alarms but often have lower TP rates as well. Thresholds on the right hand-side increase both FP and TP rates. The trade-off between TP and FP rates depends on the requirements of the specific application domains. Minimum distance optimization method assumes equal misclassification costs.

In Down syndrome prediction process sensitivity rates might be high, because we do not want to miss Down syndrome cases. However, increasing sensitivity also increases false alarms that is incorrectly detecting negative cases. Probability of false alarms corresponds to (1-specificity) and desired to have low values, because we might decrease the invasive diagnostic test.

5. EXPERIMENTS AND RESULTS

In this research, we propose a general approach to assess prenatal risk of Down syndrome by using machine learning methods. Our major aim in this study is investigating the probabilistic machine learning algorithms as a practical solution for prediction of Down syndrome using first and second trimester biochemical characteristics and screening features.

We use two different Down syndrome datasets in all our experiments. We aim to make more robust and powerful our proposed solution so we try to validate our own study and experiment result with two different datasets. Besides this, in each dataset we mention different additional points that enriches our study.

In the previous section, we detailed our methodology to conduct experiments, such as performance measures, test and training strategies, classification algorithms, dimension reduction algorithms, pre- and post-processing techniques. Besides this, during every experiment we mention them briefly as necessary. By this way we try to emphasize the importance of the results and experiments.

Throughout the whole study, we build our research over four questions. Our experiments also based on these questions to construct an integrated and consistent study. The first question focuses on benchmarking classifiers for Down syndrome prediction. For each pregnant, a data feature vector is labelled as either positive or negative. We predict this by choosing the best performed algorithm after comparing classification performance of four different classifiers. These classifiers are trained on not only original datasets but also their reduced sets.

The second question tries to find some booster ways to enhance the prediction. We answer this question according to threshold optimization and resampling strategies. We try to search some additional ways in addition to feature selection and feature extraction techniques to improve classifier performance.

The third one aims to implement probabilistic classifiers on Down syndrome data. We answer this query by applying Naive Bayes and Bayesian Networks probabilistic classifiers.

The last question aims to show the performance results apparently by benchmarking the probabilistic classifiers and other existing ones.

5.1. Experiment I: Benchmarking Non-Probabilistic Classifiers

The experiments presented in this section corresponds to the first research question: How can we construct an efficient Down syndrome prediction model?

5.1.1. Results for Dimension Reduction

Dimensionality reduction process is a pre-processing step of the proposed method. Our first dataset is composed of 31 attributes, and the second one has 24 features. As we discuss in the previous part, we conduct feature selection and feature extraction experiments and share the results in this section.

5.1.1.1. Feature Selection. We used three different well known feature selection methods; DT, SVM-RFE, Forward feature selection. Using feature selection we formed three different reduced sets for each dataset. During our classification experiments we use these reduced sets and original datasets together.

5.1.1.2. Decision Tree. We apply DT as attribute selection filter using Weka with 10 fold cross validation. For the first dataset, when it reaches the best split, the resulted tree is composed of 7 variables. These variables are A20, A23, A12, A24, A15, A3 and A5. We marked this reduced dataset as Input 1.1 for the following part of the experiments. When we apply DT feature selection filter to second dataset, the resulted tree is composed of 3

attributes including B17, B8, B16. And we labelled this reduced set as Input 1.2.

5.1.1.3. Recursive Feature Elimination with Support Vector Machine. We use Weka SVMAttrbute- Evaluator algorithm as second feature selection method. This algorithm gets a number as predefined number of variables to obtain optimal subset. We use 9 as the dimension of dataset 1 and dataset 2. With this parameters, Input 2.1 contains A23, A20, A22, A25, A26, A11, A21, A14, and A30 variables. And our second subsets involves B4, B12, B7, B6, B10, B8, B9, B2,and B1 variables. We labelled this reduced set as Input 2.2.

5.1.1.4. Forward Feature Selection. For the last feature selection method we use forward feature selection algorithm. In Weka when we use Forward direction search algorithm we get 4 dimensional subset for the first dataset which contains A20, A4, A21, and A25. We marked this subset as Input 3.1. For the second dataset, forward feature selection algorithm gives only one variable which is B4 called Input 3.2.

5.1.1.5. Feature Extraction. In feature extraction phase, we use a widely used method called PCA. PCA transforms the original datasets into reduced ones. We compose two new datasets for each of our original datasets.

5.1.1.6. Principal Component Analysis. In this work, we apply Weka PCA on full datasets where proportion of variance is defined as 0.95. PCA algorithm transforms the original Dataset 1 into a 15 dimensional input space which we mention as Input 4.1. And it generates 16 dimensional input space for our Dataset 2. We labelled this set as Input 4.2.

Table 5.1. Reduced datasets.

	Feature Selection			**Feature**
	DT	SVM-RFE	Forward	PCA
Dataset 1 (31)	Input 1.1 (7)	Input 2.1 (9)	Input 3.1 (4)	Input 4.1 (15)
Dataset 2 (24)	Input 1.2 (3)	Input 2.2 (9)	Input 3.2 (1)	Input 4.2 (16)

Numbers within the parenthesis indicate the variable number of the reduced sets. We try to summarize the feature selection and feature extraction methods results that we apply on our full datasets in Table 5.1.

5.1.2. Results for Classification

After the pre-processing phase, in classification part of this study our goal is benchmarking non-probabilistic classifiers including, k-NN, DT, SVM and MLP. After the comparison we chose the best one and then compare it with the probabilistic classifiers in the next section of the experiments. As in the previous subsection, all experiments were performed in Weka machine learning tool.

5.1.2.1. k-NN. As we mentioned before, in this study we use different k numbers as closeness factor. We apply k-NN algorithm on all of our 8 reduced sets and 2 original datasets. We use two different training and testing strategies, 10-fold cross validation and splitting(2/3 for training and 1/3 for testing purposes). Results are given in the Table 5.2. As k values we use 3 and 5 with Euclidean distance function.

Table 5.2. Classification results for k-NN algorithm.

			k=3				k=5			
	Dataset	**# features**	**Accuracy**	**Sensitivity**	**Specificity**	**FPR**	**Accuracy**	**Sensitivity**	**Specificity**	**FPR**
2/3 Split	Input 1	31	99.43%	66.67%	99.46%	0.54%	99.43%	66.67%	99.46%	0.54%
	Input 1.1	7	99.39%	-	99.39%	0.61%	99.39%	-	99.39%	0.61%
	Input 2.1	9	99.46%	75.00%	99.50%	0.50%	99.46%	75.00%	99.50%	0.50%
	Input 3.1	**4**	**99.46%**	**62.50%**	**99.57%**	**0.43%**	**99.46%**	**66.67%**	**99.53%**	**0.47%**
	Input 4.1	15	99.46%	66.67%	99.53%	0.47%	99.50%	80.00%	99.53%	0.47%
	Input 2	24	97.22%	-	97.22%	2.78%	97.22%	-	97.22%	2.78%
	Input 1.2	3	97.22%	-	97.22%	2.78%	97.22%	-	97.22%	2.78%
	Input 2.2	9	97.22%	-	97.22%	2.78%	97.22%	-	97.22%	2.78%
	Input 3.2	1	97.22%	-	97.22%	2.78%	97.22%	-	97.22%	2.78%
	Input 4.2	16	97.22%	-	97.22%	2.78%	97.22%	-	97.22%	2.78%
10-fold cross validation	Input 1	31	99.28%	60.00%	99.33%	0.67%	99.32%	77.78%	99.34%	0.66%
	Input 1.1	7	99.31%	100.00%	99.31%	0.69%	99.31%	100.00%	99.31%	0.69%
	Input 2.1	9	99.31%	58.33%	99.43%	0.57%	99.32%	60.00%	99.44%	0.56%
	Input 3.1	**4**	**99.40%**	**68.75%**	**99.52%**	**0.48%**	**99.38%**	**66.67%**	**99.50%**	**0.50%**
	Input 4.1	15	99.33%	71.43%	99.38%	0.62%	99.37%	69.57%	99.45%	0.55%
	Input 2	24	99.06%	-	99.06%	0.94%	99.06%	-	99.06%	0.94%
	Input 1.2	3	99.06%	-	99.06%	0.94%	99.06%	-	99.06%	0.94%
	Input 2.2	9	99.06%	-	99.06%	0.94%	99.06%	-	99.06%	0.94%
	Input 3.2	1	99.06%	-	99.06%	0.94%	99.06%	-	99.06%	0.94%
	Input 4.2	16	99.06%	-	99.06%	0.94%	99.06%	-	99.06%	0.94%

From the results, it can be said that k-NN classification algorithm provides feasible classification performance. According to the results, there is no difference between different k values, k=3 and k=5. They have almost the same performance measures. So we consider k=3 results for the following findings. It seems that, splitting train&test strategy has slightly better performance than 10-fold cross validation. Since this can be an effect of small data size. For splitting strategy, we can choose Input 3.1 as the best reduced set which is generated by feature selection method of forward feature selection. This experiment has maximum accuracy of 99.46% and minimum FPR of 0.43%. When we consider cross validation strategies, the best performance results come from Input 3.1 too which is formed by the same feature selection method. 99.40% accuracy and 0.48% FPR are evaluated in this experiment. It is obvious that original dataset includes some non-informative features. Input spaces which are obtain from feature selection and feature

extraction methods provide better classification performance. While the original dataset 1 consisting 31 attributes Input 3.1 include only 4 attributes. Furthermore k-NN algorithm could not work on Dataset 2 and its reduced sets. Dataset 2 has only 2 positive records out of 211. And it is very hard to predict these two positive records without sufficient training phase with sufficient positive class data. k-NN algorithm cannot be trained with these two positive records. So we mainly consider about Dataset 1 results in this part. In the second group of experiments we try to negotiate this imbalance class distribution problem.

5.1.2.2. DT. To conduct this experiment, we use Weka J48 tree as classification algorithm. We follow the same methodology in all classification performance measurement experiments. When we use DT as classification algorithm, for the splitting and the cross validation strategies the best performance results are belong to Input 3.1 as in the kNN method, 99.50% Accuracy, 0.40% FPR and 99.49% Accuracy, 0.40 FPR respectively as shown in Table 5.3. DT confirms the positive effect of dimension reduction methods effect. Reduced sets have better performance results than the original ones. Since in this phase, we can say that feature selection method outperforms the feature extraction method. And also feature extraction method performs worse than original dataset. We can say that feature extraction technique fails with DT classification algorithm for our datasets. Similar to the previous experiments DT also fails for Dataset 2. It could not classify any true positive outcome out of 2 true positive records.

Table 5.3. Classification results for DT algorithm.

	Dataset	# features	Accuracy	Sensitivity	Specificity	FPR
2/3 Split	Input 1	31	99.46%	57.14%	99.68%	0.32%
	Input 1.1	7	99.46%	57.14%	99.68%	0.32%
	Input 2.1	9	99.43%	54.55%	99.60%	0.40%
	Input 3.1	**4**	**99.50%**	**66.67%**	**99.60%**	**0.40%**
	Input 4.1	15	99.43%	60.00%	99.50%	0.50%
	Input 2	24	97.22%	-	97.22%	2.78%
	Input 1.2	3	97.22%	-	97.22%	2.78%
	Input 2.2	9	97.22%	-	97.22%	2.78%
	Input 3.2	1	97.22%	-	97.22%	2.78%
	Input 4.2	16	97.22%	-	97.22%	2.78%
10-fold cross validation	Input 1	31	99.35%	61.76%	99.51%	0.49%
	Input 1.1	7	99.40%	68.75%	99.52%	0.48%
	Input 2.1	9	99.43%	66.67%	99.60%	0.40%
	Input 3.1	**4**	**99.49%**	**75.68%**	**99.60%**	**0.40%**
	Input 4.1	15	99.11%	89.66%	99.35%	0.65%
	Input 2	24	99.06%	-	99.06%	0.94%
	Input 1.2	3	99.06%	-	99.06%	0.94%
	Input 2.2	9	99.06%	-	99.06%	0.94%
	Input 3.2	1	99.06%	-	99.06%	0.94%
	Input 4.2	16	99.06%	-	99.06%	0.94%

5.1.2.3. MLP. MLP is a supervised classifier which provides high performance when acquiring hidden knowledge. In our MLP experiments, we use 1 hidden layer, 10 hidden units, 20 epochs with a learning rate of 0.3 and momentum as 0.2.

From the results, as shown in Table 5.4 we can see that MLP provides high classification performance. For the splitting strategy Input 1 provides 99.57% accuracy and 0.32% FPR as the best performance. And for the cross validation strategy Input 1.1 produces 99.46% accuracy with 0.44 FPR. For splitting strategy original dataset provides better results than reduced ones. And also feature extraction and feature selection methods have almost the same performance results.

While previous classification algorithms could not classify any true positive outcome for Dataset 2, MLP works over Dataset 2 and predict a true positive outcome out of 2. MLP performs well on the original dataset 2.

Table 5.4. Classification results for MLP algorithm.

	Dataset	# features	Accuracy	Sensitivity	Specificity	FPR
2/3 Split	**Input 1**	**31**	**99.57%**	**72.73%**	**99.68%**	**0.32%**
	Input 1.1	7	99.53%	75.00%	99.61%	0.39%
	Input 2.1	9	99.46%	62.50%	99.57%	0.43%
	Input 3.1	4	99.50%	66.67%	99.60%	0.40%
	Input 4.1	15	99.53%	70.00%	99.60%	0.36%
	Input 2	24	97.22%	-	97.22%	2.78%
	Input 1.2	3	97.22%	-	97.22%	2.78%
	Input 2.2	9	97.22%	-	97.22%	2.78%
	Input 3.2	1	97.22%	-	97.22%	2.78%
	Input 4.2	16	97.22%	-	97.22%	2.78%
10-fold cross validation	Input 1	31	99.38%	61.90%	99.57%	0.43%
	Input 1.1	**7**	**99.46%**	**75.76%**	**99.56%**	**0.44%**
	Input 2.1	9	99.42%	68.57%	99.55%	0.45%
	Input 3.1	4	99.42%	67.57%	99.56%	0.44%
	Input 4.1	15	99.38%	63.16%	99.55%	0.45%
	Input 2	**24**	**99.06%**	**50.00%**	**99.53%**	**0.47%**
	Input 1.2	3	99.06%	-	99.06%	0.94%
	Input 2.2	9	99.06%	-	99.06%	0.94%
	Input 3.2	1	99.06%	-	99.06%	0.94%
	Input 4.2	16	98.12%	0.00%	99.05%	0.95%

5.1.2.4. SVM. We apply SVM as the last non-probabilistic classification algorithm. SVM is a well known discriminant analysis method. We train SVM with polynomial kernel function. We use cost parameter as 1 and tolerance parameter as 0.0010. Classification results for SVM with these parameters are given in Table 5.5.

SVM produces high classification performance on original datasets and their reduced sets. It performs almost same with feature selection and feature extraction methods. And the full dataset performs as well as its reduced subsets with feature selection methods. Since reduced sets shorten the execution time of the algorithm. For splitting strategy three of the feature selection methods performs almost the same and are better than the original dataset. We choose Input 1.1 as the best one with regarding the AUC measure. SVM has 99.53% accuracy and 0.39% FPR with Input 1.1. When we consider 10-fold cross validation strategies SVM has the best results with Input 1.

Similar to MLP algorithm, SVM also works for Dataset 2. It can classify one of the true positive outcomes in Dataset 2 and its reduced sets except Dataset 3.2 which is formed by forward feature selection algorithm.

Table 5.5. Classification results for SVM algorithm.

	Dataset	# features	Accuracy	Sensitivity	Specificity	FPR
2/3 Split	Input 1	31	99.46%	62.50%	99.57%	0.43%
	Input1.1	**7**	**99.53%**	**75.00%**	**99.61%**	**0.39%**
	Input 2.1	9	99.53%	75.00%	99.61%	0.39%
	Input 3.1	4	99.53%	75.00%	99.61%	0.39%
	Input 4.1	15	99.50%	71.43%	99.57%	0.43%
	Input 2	24	97.22%	-	97.22%	2.78%
	Input1.2	3	97.22%	-	97.22%	2.78%
	Input 2.2	9	97.22%	-	97.22%	2.78%
	Input 3.2	1	97.22%	-	97.22%	2.78%
	Input 4.2	16	97.22%	-	97.22%	2.78%
10-fold cross validation	**Input 1**	**31**	**99.44%**	**77.78%**	**99.51%**	**0.49%**
	Input 1.1	7	99.43%	75.00%	99.51%	0.49%
	Input 2.1	9	99.38%	69.23%	99.47%	0.53%
	Input 3.1	4	99.40%	73.08%	99.49%	0.51%
	Input 4.1	15	99.39%	72.0%	99.48%	0.52%
	Input 2	24	98.53%	100.00%	99.53%	0.47%
	Input 1.2	3	98.59%	33.33%	99.52%	0.48%
	Input 2.2	**9**	**99.06%**	**50.00%**	**99.53%**	**0.47%**
	Input 3.2	1	99.06%	-	99.06%	0.94%
	Input 4.2	16	98.59%	33.33%	99.52%	0.48%

5.1.2.5. Classifier Evaluation. Up to this point we apply four different non-probabilistic classification algorithms with different train and test strategies using different input sets.

While determining the best techniques to use as main non-probabilistic classifier in the rest of the experiments, we should consider each performance measures. We mostly consider minimizing FPR at the same time with maximum accuracy. One of the main goal of our study is minimizing invasive diagnostic tests to ensure the DS positive outcome case. To prevent unnecessary invasive operations we might minimize the FPR. As a result, we take FPR and accuracy into account when choosing the optimal classifiers.

Table 5.6. Summary of classifier performance.

Classifier	Test/Train M.	Dataset	Accuracy	FPR
k-NN	Splitting	Input 3.1	99.46%	0.43%
k-NN	Cross Val.	Input 3.1	99.40%	0.48%
SVM	**Splitting**	**Input 1.1**	**99.53%**	**0.39%**
SVM	Cross Val.	Input 1	99.44%	0.49%
DT	Splitting	Input 3.1	99.50%	0.40%
DT	Cross Val.	Input 3.1	99.49%	0.40%
MLP	**Splitting**	**Input 1**	**99.57%**	**0.32%**
MLP	Cross Val.	Input 1.1	99.46%	0.44%
MLP	Cross Val.	Input 2	99.06%	0.47%
SVM	Cross Val.	Input 2.2	99.06%	0.47%

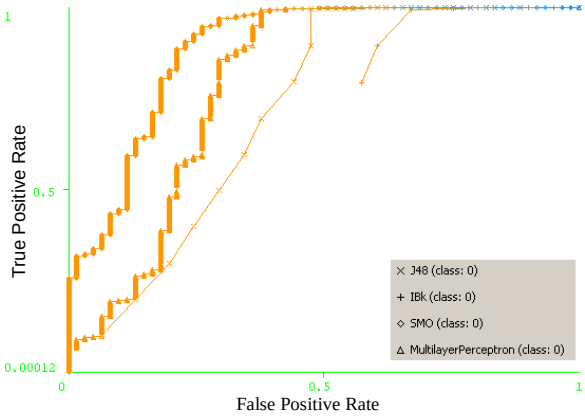
Table 5.7. Performance summary of all classifiers.

Dataset	Classifier	Accuracy	FPR	AUC
Input 1	SVM	99.44%	0.49%	0.887
Input 1	MLP	99.38%	0.43%	0.792
Input 1	DT	99.35%	0.49%	0.717
Input 1	kNN	99.28%	0.67%	0.652
Input 1.1	MLP	99.46%	0.44%	0.915
Input 1.1	SVM	99.43%	0.49%	0.803
Input 1.1	DT	99.40%	0.48%	0.717
Input 1.1	kNN	99.31%	0.69%	0.667
Input 2.1	MLP	99.42%	0.45%	0.881
Input 2.1	SVM	99.38%	0.53%	0.791
Input 2.1	DT	99.43%	0.40%	0.717
Input 2.1	kNN	99.31%	0.57%	0.729
Input 3.1	MLP	99.42%	0.44%	0.908
Input 3.1	SVM	99.40%	0.51%	0.811
Input 3.1	DT	99.49%	0.40%	0.729
Input 3.1	kNN	99.40%	0.48%	0.726
Input 4.1	SVM	99.39%	0.52%	0.881
Input 4.1	MLP	99.38%	0.45%	0.822
Input 4.1	DT	99.11%	0.65%	0.78
Input 4.1	kNN	99.33%	0.62%	0.676

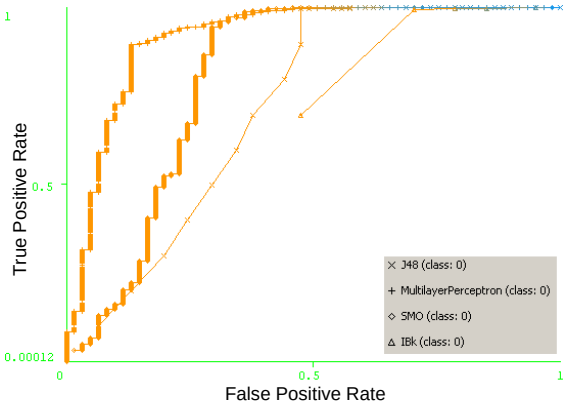
Table 5.6 shows the best performance measures of each non-probabilistic classifier. According to Table 5.6. none of the classifiers have a significantly better performance than others. None of them can be selected as the best since none of them producing the best

performance on all inputs. We see that, on the original dataset 1, MLP provides the highest accuracy with minimum FPR and for Input 1.1, SVM reaches the second highest accuracy among all classifiers with the second lowest FPR. Actually all of the results are close each other. Since we might choose an optimal algorithm for further comparisons with probabilistic classifiers performances. From this point of view, we prefer to use MLP and SVM in our next experiments.

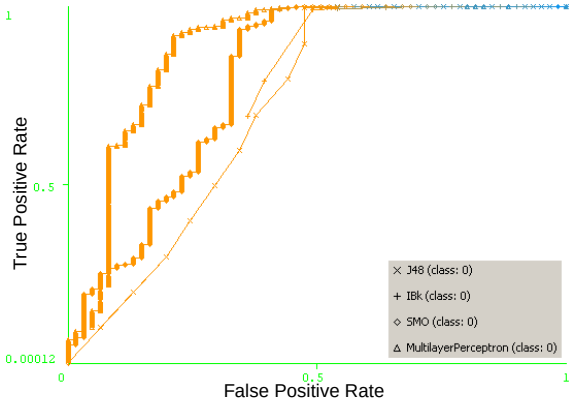
To display the performance measures obviously and explain why we choose MLP and SVM, we use ROC curves of all the classifiers performances. Table 5.7 shows the summary of all classifiers performance measures by each different input and Figure 5.1 shows ROC analysis of four classifiers for five input sets of Dataset 1. By ROC analysis we provide an obvious benchmarking demonstration.



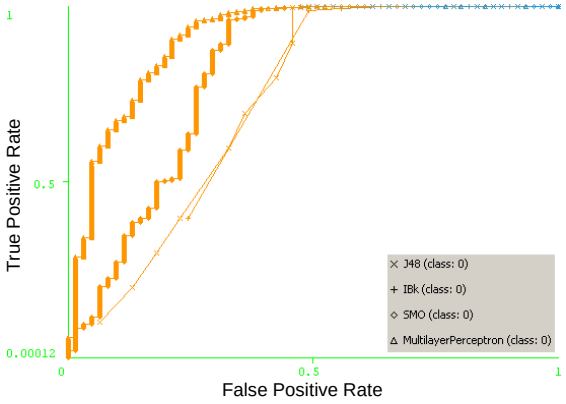
(a) ROC curve with Input 1.



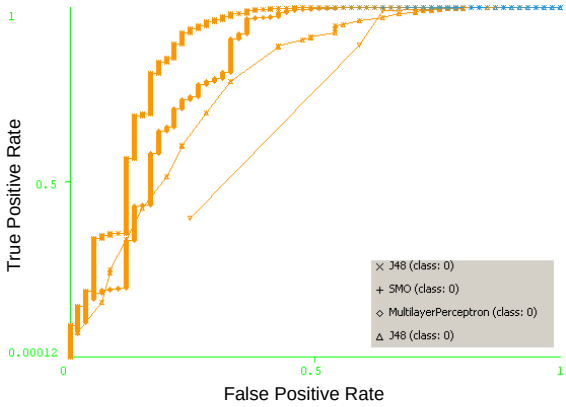
(b) ROC curve with Input 1.1.



(c) ROC curve with Input 2.1.



(d) ROC curve with Input 3.1.



(e) ROC curve with Input 4.1.

Figure 5.1. ROC analysis representation for five different inputs.

Table 5.8 shows our optimal classifiers' confusion matrix for splitting train&test strategy. Right hand side table generated from SVM classification over Input 1.1 and

splitting strategy. And the other table generated by MLP classification method over Input 1 and splitting strategy. All of the above performance measures calculated from confusion matrix tables. We choose MLP and SVM as our optimal non-probabilistic classifiers.

Table 5.8. Confusion matrix for SVM & MLP classifiers.

(a) MLP

		Predicted	
		Positive	Negative
Actual Case	Positive	8	9
	Negative	3	2773

(b) SVM

		Predicted	
		Positive	Negative
Actual Case	Positive	6	11
	Negative	2	2774

5.2. Experiment II: Outperform the Prediction Results

In this section we have some experiments corresponding to the second research question: How can we enhance the methodologies to improve the prediction performance?

Datasets that we use in our study contain fewer samples with positive outcomes. Any classifier built on these datasets has much more information to identify negative cases compared to positive DS cases. Therefore, DS prediction is handled as a typical case of learning from imbalanced data problem. In this part of the experiments, we investigate the effects of resampling methods in prediction performance in case of imbalanced distribution. As we see in the first part of the experiment two of the classifiers could not detect any positive outcome out of two positive instances. And as a second method to negotiate the imbalanced class distribution problem we apply threshold optimization technique.

In this phase, we also investigate categorical variable transformation effect by implementing two different encoding techniques, binary encoding and frequency based encoding.

There are two main resampling strategies, over-sampling that replicates instances from minority class [43] and under-sampling where some of the instances in the minority class is removed [42].

In Experiment I, we have compared various classifiers for DS prediction and show that SVM produces slightly better predictive performance with original dataset, Input 1. Therefore, we apply SVM in order to investigate the effect of resampling strategies.

5.2.1. Results for Resampling

We implement oversampling method for both original datasets, on the other hand apply undersampling method for only Dataset 1. Because there are only 2 positive instances in Dataset 2 and this structure is not suitable for undersampling method.

Table 5.9. Distribution of class samples and prediction results after over sampling the Dataset 2.

Dataset No	1	**2**	3	4	5	6	7	8	9	10
# Positive Samples	2	**15**	30	60	120	180	240	300	360	420
# Negative Samples	211	**211**	211	211	211	211	211	211	211	211
True Positive Rate (%)	50.00	**100.00**	96.77	98.36	100.00	100.00	99.59	99.67	99.72	99.76
False Positive Rate (%)	0.47	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Positive/Negative (%)	0.95	**7.11**	14.22	28.44	56.87	85.31	113.74	142.18	170.62	199.05

For over sampling, we have constructed ten training sets by replicating the positive instances while keeping the number of negative instances constant. Table 5.9 shows the exact instance numbers according to different classes for Dataset 2. According to results, TPR increases its maximum value at the second fold of oversampling with minimum FPR. Therefore we choose the second sample as the optimum one with 100% TPR and 0% FPR. This can be interpreted as increasing the number of positive samples raise the number of positive predictions.

Table 5.10. Distribution of class samples and prediction results after undersampling the Dataset 1.

Dataset No	1	2	**3**	4	5	6	7	8	9	10
# Positive Samples	61	61	**61**	61	61	61	61	61	61	61
# Negative Samples	8155	1000	**500**	250	110	70	55	40	35	30
True Positive Rate (%)	69.2	86.0	**88.89**	87.23	87.04	86.21	85.25	85.71	87.88	87.30
False Positive Rate (%)	0.53	2.36	**4.07**	7.58	11.97	15.07	16.36	18.42	10.00	21.43
Positive/Negative(%)	0.75	6.10	**12.20**	24.40	55.45	87.14	110.91	152.50	174.29	203.33

Table 5.11. Distribution of class samples and prediction results after oversampling the Dataset 1.

Dataset No	1	2	**3**	4	5	6	7	8	9	10
# Positive Samples	61	500	**1000**	2000	4500	7000	9000	12500	14000	16000
# Negative Samples	8155	8155	**8155**	8155	8155	8155	8155	8155	8155	8155
True Positive Rate (%)	69.2	87.8	**88.9**	89.38	89.07	89.78	90.17	91.35	91.62	92.22
False Positive Rate (%)	0.53	2.45	**3.84**	6.10	9.83	11.36	11.75	10.48	10.79	10.43
Positive/Negative(%)	0.75	6.13	**12.26**	24.52	55.18	85.84	110.36	153.28	171.67	196.20

Table 5.10 and Table 5.11 represent the distribution of the Dataset 1 and prediction results in terms of TPR and FPR for under sampling and over sampling respectively. Similar to the over sampling of Dataset 2 experiment, both TPR and FPR increase at each fold of resampling up to a certain level.

The trade-off between the TPR and FPR can be adjusted by changing the ratio of classes. Optimum TPR and FPR pair can also be obtained as explained in Section 4.8. These corresponds to (88.89%, 4.07%) and (88.98%, 3.84) for undersampling and oversampling respectively.

5.2.2. Results for Threshold Optimization

As the second method to solve the imbalanced class problem, we calculate the TPR and FPR values by varying the decision thresholds in range of [0:0.1:1]. The resulting set of (TPR, FPR) pairs are given in Table 5.12 and Table 5.13.

Classification with the default decision threshold, 0.5, produce 77.78% TPR and 0.49% FPR for Dataset 1. And default threshold produces 100% TPR and 0.47% FPR for Dataset 2. According to the below results we can say that the default threshold is optimal one. Choosing a different threshold rather than default one could not produce better results. For our datasets threshold optimization could not work well.

Table 5.12. Prediction results depending on variation of decision threshold – Dataset 1.

Decision Threshold	0.1	0.2	0.3	0.4	**0.5**	0.6	0.7	0.8	0.9
True Positive Rate (%)	56.36	67.50	69.70	75.00	**77.78**	77.27	77.78	76.92	80.00
False Positive Rate (%)	0.37	0.42	0.46	0.49	**0.49**	0.54	0.57	0.62	0.65

Table 5.13. Prediction results depending on variation of decision threshold – Dataset 2.

Decision Threshold	0.1	0.2	0.3	0.4	**0.5**	0.6	0.7	0.8	0.9
True Positive Rate (%)	16.67	25.00	50.00	100.00	**100.00**	100.00	100.00	-	-
False Positive Rate (%)	0.48	0.48	0.47	0.47	**0.47**	0.47	0.47	0.94	0.94

As we consider both resampling and threshold optimization techniques, we prefer to use resampling strategies to improve the classification performances. Resampling methods improve our performance, since, we cannot say it for threshold optimization.

5.2.3. Results for Transformation of Categorical Variables

In addition to resampling and threshold optimization techniques we investigate nominal variables effect by transforming them into numeric variables.

In our second set of experiments we use SVM as classification algorithm and in distance based classifiers, such as SVM, performance depends on accurate transformation of categorical variables into numeric data.

We use two different transformation techniques during these experiments. First one is a nonsupervised method, common binary encoding approach and the second one is a supervised method, frequency based encoding technique .

Dataset 1 includes two categorical variables: A7 and A10 with 7 and 2 categories respectively. And Dataset 2 includes three categorical variables: B4, B12, and B14 with 2 categories.

5.2.3.1. Binary Encoding. For a particular categorical variable including N categories, each category represented by a sequence of N bits. The i*th* bit corresponding to original category is set to 1 and the others are set to 0. When binary encoding is applied, the categories 1,2,..,7 correspond to 0000001, 0000010,..., 1000000 respectively. In this case Euclidean distance between each category is equal, however, this may not be the actual case.

5.2.3.2. Frequency Based Encoding. Numerical values are derived from the relative frequencies of categorical codes among both positive and negative classes. The basic idea behind this transformation is to reflect the effect of categorical code on prediction [99]. Hence, the numerical value of a categorical code is defined as the difference between frequencies in positive and negative classes in the range of [-1,1]. The categories 1,2,..,7 correspond to -0.0424, -0.8168,...,-0.0096 as a result of frequency transformation as shown in Table 5.14.

Table 5.14. Distribution of categorical variables among both positive and negative classes.

(a) A7 Feature

A7 – Categories	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Positive	0.0002	0.0063	0.0005	0.0004	0	0	0
Negative	0.426	0.8231	0.0545	0.0517	0.0099	0.0011	0.0096

(b) A10 Feature

A10 – Categories	**Y**	**N**
Positive	0.0001	0.0202
Negative	0.0073	0.9724

Table 5.14 represents the distribution of the categorical variables among both positive and negative classes for Dataset 1.

By using the frequencies in Table 5.14 we transform categorical variables into numeric data. We demonstrate a transformation example in Table 5.15 for both binary encoding and frequency based encoding.

Table 5.15. Example transformation of A7 feature including 7 categories.

Original category code	Binary encoding	Frequency based encoding
1	0000001	-0.0424
2	0000010	-0.8168
3	0000100	-0.0540
4	0001000	-0.0514
5	0010000	-0.0099
6	0100000	-0.0011
7	1000000	-0.0096

We analyse the transformation effect using SVM classification algorithm. We apply both binary and frequency based encoding for Dataset 1, however, apply only binary encoding method for dataset 2. Because for three of the variables positive class frequency is zero. We show the results in Table 5.16. Transformation of categorical variables does not have a positive effect on prediction accuracy.

Table 5.16. Comparison of transformation methods for categorical variables.

	Input 1 + SVM	
	Acc	FPR
Binary	99.44%	0.49%
Frequency	99.43%	0.50%
Difference	-0.01%	-2.43%

5.3. Experiment III: Implementing Probabilistic Classifiers

Experiments presented in this section correspond to the third research question: Is it possible to predict DS with a probabilistic classifier accurately?

To investigate the performance of probabilistic approaches, we apply Naive Bayes and Bayesian Networks as probabilistic classifiers. We mainly focus on classification methods based on probability theory. Bayes theorem plays a critical role in probabilistic learning and classification.

5.3.1. Naive Bayes

Naive Bayes is a simple probabilistic classifier based on Bayes theorem. We use NaiveBayes algorithm within Weka to implement these experiments. In addition we use supervised discretization to convert numeric attributes to nominal ones.

We use two different train and test strategies during probabilistic classifier experiments. And we use both original datasets and their reduced sets. NB produces acceptable performance results. The best result is produced with transformed input by feature extraction for both train and test strategies. Actually we can say reduced sets perform better than the original datasets for NB. Similar to kNN and DT, NB also could not work for Dataset 2. It could not predict any positive classes. In contrast to the previous classifiers, NB is very fast and has very short execution time.

Feature extraction pre-processing with NB produces 99.46% accuracy with 0.47 FPR for splitting strategy as shown in the Table 5.17. And for 10 fold cross validation they are 99.25% and 0.62%. These results are satisfactory when we consider other classification methods. Therefore we show that our first probabilistic classifier performs as good as non-probabilistic ones.

Table 5.17. Classification results for NB algorithm.

	Dataset	# features	Accuracy	Sensitivity	Specificity	FPR
2/3 Split	Input 1	31	97.21%	11.39%	99.71%	0.29%
	Input 1.1	7	99.18%	31.25%	99.57%	0.43%
	Input 2.1	9	97.99%	14.55%	99.67%	0.33%
	Input 3.1	4	99.28%	42.11%	99.68%	0.32%
	Input 4.1	**15**	**99.46%**	**66.67%**	**99.53%**	**0.47%**
	Input 2	24	97.22%	-	97.22%	2.78%
	Input 1.2	3	97.22%	-	97.22%	2.78%
	Input 2.2	9	97.22%	-	97.22%	2.78%
	Input 3.2	1	97.22%	-	97.22%	2.78%
	Input 4.2	105	97.22%	-	97.22%	2.78%
10-fold cross validation	Input 1	31	95.41%	11.61%	99.77%	0.23%
	Input 1.1	7	99.09%	37.04%	99.50%	0.50%
	Input 2.1	9	97.68%	17.50%	99.68%	0.32%
	Input 3.1	4	99.03%	38.82%	99.66%	0.34%
	Input 4.1	**15**	**99.25%**	**47.62%**	**99.38%**	**0.62%**
	Input 2	24	99.06%	-	99.06%	0.94%
	Input 1.2	3	99.06%	-	99.06%	0.94%
	Input 2.2	9	99.06%	-	99.06%	0.94%
	Input 3.2	1	99.06%	-	99.06%	0.94%
	Input 4.2	16	99.06%	-	99.06%	0.94%

5.3.2. Bayesian Networks

As our second probabilistic classifier we apply Bayesian Networks which is also applying Bayes theorem using directed acyclic graphical model. We use Weka BayesNet classifier with BMAEstimator to estimate conditional probability tables of the network. And as the search algorithm we use hill climbing algorithm.

At the first look, we can say that BN performs slightly better than NB. In contrast to NB, BN performs better with DT feature selection algorithm. According to the Table 5.18 the best results belong to feature selection method. It performs 99.50% accuracy with 0.43% FPR when we use splitting strategy, and for the cross validation strategy it has 99.39% accuracy with 0.54% FPR. Reduced sets perform better than the original datasets. Our second probabilistic classifier also performs as well as non-probabilistic methods.

Table 5.18. Classification results for BN algorithm.

	Dataset	# features	Accuracy	Sensitivity	Specificity	FPR
2/3 Split	Input 1	31	98.14%	15.69%	99.67%	0.33%
	Input 1.1	**7**	**99.50%**	**71.43%**	**99.57%**	**0.43%**
	Input 2.1	9	98.14%	15.69%	99.67%	0.33%
	Input 3.1	4	99.50%	71.43%	99.57%	0.43%
	Input 4.1	15	99.39%	50.00%	99.64%	0.36%
	Input 2	24	99.25%	41.67%	99.75%	0.25%
	Input 1.2	3	97.22%	-	97.22%	2.78%
	Input 2.2	9	97.22%	-	97.22%	2.78%
	Input 3.2	1	97.22%	-	97.22%	2.78%
	Input 4.2	16	97.22%	-	97.22%	2.78%
10-fold cross validation	Input 1	31	97.35%	15.72%	99.69%	0.31%
	Input 1.1	**7**	**99.39%**	**73.91%**	**99.46%**	**0.54%**
	Input 2.1	9	97.35%	15.72%	99.69%	0.31%
	Input 3.1	4	99.37%	68.00%	99.46%	0.54%
	Input 4.1	15	99.16%	44.29%	99.63%	0.37%
	Input 2	24	99.06%	-	99.06%	0.94%
	Input 1.2	3	99.06%	-	99.06%	0.94%
	Input 2.2	9	99.06%	-	99.06%	0.94%
	Input 3.2	1	99.06%	-	99.06%	0.94%
	Input 4.2	16	99.06%	-	99.06%	0.94%

5.3.2.1. Classifier Evaluation. In this step we apply two different probabilistic classification algorithms with different train and test strategies using different input sets.

Table 5.19. Summary of classifier performance.

Classifier	Test/Train M.	Dataset	Accuracy	FPR
NB	Splitting	Input 4.1	99.46%	0.47%
NB	Cross Val.	Input 4.1	99.25%	0.62%
BN	Splitting	Input 1.1	99.50%	0.43%
BN	Cross Val.	Input 1.1	99.39%	0.54%

As shown in the Table 5.19 performance measures are very similar for both probabilistic methods. These results are the best performance measures for each classifier. For further comparison experiments we use both NB and BN as probabilistic algorithms.

5.4. Experiment IV: Probabilistic Classifiers vs. Non-Probabilistic Classifiers

As the last group of experiments we try to answer the last research question: Does probabilistic method in determining DS outperform existing methods?

In the previous sections we benchmark four different non-probabilistic classifiers and two probabilistic algorithms within each group. We have chosen two methods as our optimal non-probabilistic classifiers, and have chosen two algorithms as probabilistic classifiers. In this section, we compare these four methods to demonstrate the "probabilistic classifiers vs. non-probabilistic classifiers" scenario. Our optimal non probabilistic classifiers were MLP and SVM, and probabilistic classifiers were NB and BN. Table 5.20 shows these classifiers' best performance measures.

Table 5.20. Performance summary of four classifiers.

Dataset	Classifier	Accuracy	FPR	AUC
Input 4.1	NB	99.25%	0.62%	0.896
Input 1.1	BN	99.39%	0.54%	0.873
Input 1.1	MLP	99.46%	0.44%	0.915
Input 1	SVM	99.44%	0.49%	0.887

Every classifier performs its best performance measures with different pre-processing methods. NB performs best with PCA feature extraction, BN and MLP perform best with DT feature selection, and SVM performs best without feature extraction and feature selection methods. As shown in the Table 5.20, it is obvious that none of the classifier can be selected as the best as none of them producing the best performance on all inputs. As a result, we aim to demonstrate the probabilistic classifiers' performance within our study. So we choose Input 4.1 which is derived with PCA feature extraction method and benchmark four different classifiers using it.

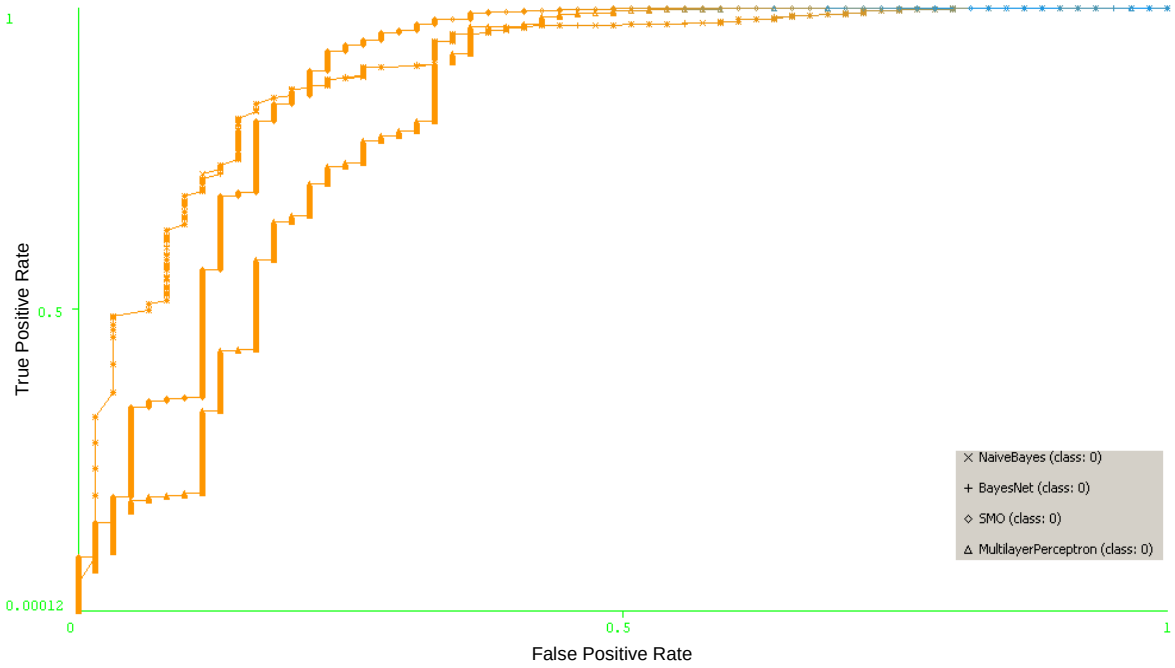


Figure 5.2. ROC analysis representation for four different classifiers.

Figure 5.2 shows the ROC analysis representation using Input 4.1 for four different classification algorithms. For Input 4.1 BN and NB have almost the same performance measures and slightly better than MLP and SVM. This difference is shown in ROC analysis and also Table 5.21 shows the exact AUC measures.

Table 5.21. Performance results for Input 4.1.

Dataset	**Classifier**	**Accuracy**	**FPR**	**AUC**
Input 4.1	NB	99.25%	0.62%	0.896
Input 4.1	BN	99.16%	0.37%	0.897
Input 4.1	MLP	99.38%	0.45%	0.822
Input 4.1	SVM	99.39%	0.52%	0.881

All the experiments considered, we can say that probabilistic classifiers perform as well as non-probabilistic ones. Even we prove that probabilistic classifiers can outperform non-probabilistic methods.

5.5. Discussion

In medical domain, the most sensitive method for Trisomy 21 screening was introduced as the combination of maternal age, serum screening for PAPP-A, f β-hCG and fetal NT with 90% detection rate and 5% FPR [24]. And also the studies that we obtained our datasets had similar findings in terms of first and second trimester screening for Trisomies 21. The first study provided 65% DR and 5% FPR [55]. In this work commercial risk assessment software is used and the 1/270 risk level was used as the threshold value considering positive Down syndrome outcome.

The second research that we gathered our Down syndrome dataset had similar findings. They reported 70% accuracy at 8.8 FPR. Both of these studies used the same set of features which are widely used characteristics in clinical domain. Within this study risk analysis for Trisomies 21 was made by the commercial PRISCA software considering the cut-off value 1/250.

The results of the first group of experiments showed that a reliable dimensionality reduction pre-processing was possible with machine learning techniques. We did not use expert knowledge while choosing our input features. Since, we employ dimensionality reduction methods to eliminate irrelevant features within whole data, and got the same features within these studies. When we employed DT feature selection technical, it selected maternal age, f β-hCG, PAPP-A and fetal NT as the most informative features. These characteristics are the most widely used ones in clinical domain and with our proposed solution we can choose these features without any expert knowledge.

The proposed method in machine learning domain also provided acceptable prediction performance. Experiments I and III show that classification algorithms produced better results than previous works. Both non-probabilistic classifiers and probabilistic classifiers may have up to 100% sensitivity with less than 1% FPR with further machine learning techniques. For example BN classifier has 73.1% sensitivity with 0.54 FPR which provides 99.39% accuracy with DT feature selection. And SVM provides 77.78 sensitivity with 0.49 FPR that means 99.44 accuracy.

The model has been validated in a prospective manner and results supported the classification performance of the proposed system. It is expected that the presented Down

syndrome prediction model will provide useful information for decision-making on deciding further invasive diagnostic test.

6. CONCLUSIONS

We prefer to categorize our conclusion as overall summary, medical perspective of our study and future research directions.

6.1. Overall Summary

In this research, we present the machine learning approach as a solution to the prenatal risk assessment of Down syndrome. We have concentrated on probabilistic approaches for predicting the DS fetal abnormality. Our major goal was evaluating prenatal risk of DS with a non-invasive method, especially with a machine learning method applying probability theorem. To accomplish this, first, we needed to understand the existing methods, the difficulty faced in those approaches. Then, we designed our experiments in an iterative manner to match the clinical requirements to machine learning problems.

In this study we tried to answer four research questions. The experimental design was mainly hold in terms of these questions. First of them was: How can we construct an efficient Down syndrome prediction model?

To answer first question we formalized the problem as supervised binary classification problem. From a machine learning perspective, the imbalanced class ratio of positive and negative samples entailed the investigation of prediction performance in terms of TPR and FPR rather than single accuracy measure. Also, sensitivity and specificity are the common performance measures in the medical literature. Therefore, performance criteria was based on the ROC analysis in our experiments.

To enrich and self validation we used two different DS databases from different sources. Generally we conducted our experiments for each dataset with the same way. Moreover, the number of instances of these datasets were significantly different. We tried to design different experimental scenarios to turn in favour of enriching the study.

As a pre-process phase, we investigate how to reduce our feature space in order to prevent from increasing time and space complexity and also eliminate redundant variables and hence to increase prediction performance. We compare performance of both feature selection algorithms to get the most informative features and feature extraction algorithm to map our sample space into lower dimensional space. For this comparison, different techniques are applied on original datasets that produces a number of reduced input sets. As comparison criterion, we use these reduced sets and also original datasets with four different non-probabilistic classifiers. After a comparative analysis of diverse classifiers, we decided SVM and MLP to be the best fitting algorithms for DS prediction problem. We choose these classifiers for further experiments, for instance, benchmarking with probabilistic classifiers.

In the second part of the study, we have designed experiments to improve the prediction performance by using methodological enhancements to standard machine learning algorithms. First, we compared resampling methods, and decision threshold optimization in order to handle the imbalance class problem. Experiments show that resampling the training data can provide better results. Second, we analysed transformation of categorical variables effect by using two different encoding techniques. We could not get a considerable improvement applying encoding techniques.

The next research direction was to predict the DS abnormality by probabilistic classifiers. We have used Naive Bayes and Bayesian Networks as probabilistic classifiers which apply Bayes theorem. The results of the experiments revealed that both probabilistic classifiers perform similar, with closed performance measures which are acceptable with a high prediction performance in terms of AUC measure.

In the last part, we have compared probabilistic classifiers performance and non-probabilistic ones. As non-probabilistic classifiers, we used algorithms that we choose as optimal in first stage of the work, SVM and MLP. And as probabilistic classifiers we apply NB and BN. According to the held experiments probabilistic classifiers have slightly better performance results than non-probabilistic methods in terms of AUC measure.

To conclude, our study presented the potential of machine learning algorithms in increasing the prediction rates in DS abnormalities as a non-invasive approach. Based on

our findings, we can advice using probabilistic classifiers such as NB and BN for prenatal risk assessment of DS.

6.2. The Clinical Perspective

We mainly concentrated on reducing invasive operations to diagnose Down syndrome utilizing machine learning techniques. Our proposed method especially uses first and second trimester biochemical characteristics. As we revealed in the experiments part our method selects the most informative features which are recently most widely used characteristics by obstetrician in medical domain. Maternal age, f β-hCG, PAPP-A and fetal NT characteristics were selected as the most interesting features in our feature selection pre-processing step. By this way, we can say that this validates our method with considering expert knowledge in medical domain.

There are some recent studies to try to identify new features effect that characterizing Down syndrome outcome like ductus venosus doppler measurements and tricuspid regurgitation. Our method can be employed in these studies to evaluate new features effect that causes Down syndrome. When the input space is increased by obtaining more biochemical or screening measurements there may be selected new features as characterizing positive Down syndrome outcome.

Our proposed model designed to be an alternative way to predict Down syndrome cases before invasive operations in an early phase as possible. We utilized machine learning techniques to accomplish this. We do not consider it as the replacement of certain invasive diagnostic tests. To the best of our knowledge, this kind of work using probabilistic classification techniques is novel probabilistic model in the medical domain.

6.3. Future Research Directions

We conducted our experiments on two different Down syndrome datasets and analyze results for each separate input set. There can be obtained one more Down

syndrome dataset with similar features. After selecting same features from these three datasets, these features can be normalized within each group in terms of MoM metrics. Then the proposed method can be trained using first dataset, tested with second dataset and validated with the third dataset. By this way the proposed method can be evaluated more accurately in terms of practical usage by physicians.

We have mostly concentrated on probabilistic methods including Bayesian Networks. We used BMAEstimator algorithm to estimate conditional probability tables of the network. However, by physicians expert knowledge may be included and an optimal network structure that represents the Down syndrome can be further investigated.

REFERENCES

1. Down, L. J., "Observations on an Ethnic Classification of Idiots", London Hospital Clinical Lectures and Reports, No. 3, pp. 259-262, 1866.

2. Brambati, B., M. C., Macintosh, B., Teisner, S., Maguiness, K., Shrimanker, A., Lanzani, "Low Maternal Serum Levels of Pregnancy Associated Plasma Protein A (PAPP-A) in the First Trimester in Association with Abnormal Fetal Karyotype", *British Journal of Obstetrics and Gynaecology*, Vol. 100, pp. 324-330, 1993.

3. Spencer, K., J. N., Macri, D. A., Aitken, J. M., Connor, "Free Beta-hCG as First Trimester Marker for Fetal Trisomy", *Lancet*, Vol. 339, p. 1480, 1992.

4. Snijders, R. J., P., Noble, N., Sebire, A., Souka, K. H., Nicolaides, "UK Multicentre Project on Assessment of Risk of Trisomy 21 by Maternal Age and Fetal Nuchal Translucency Thickness at 10-14 Weeks of Gestation", *Lancet*, Vol. 352, pp.343-349, 1998.

5. Nicolaides, K. H., K., Spencer, K., Avgidou, S., Faiola, O., Falcon, "Multicenter Study of First Trimester Screening for Trisomy 21 in 75821 Pregnancies: Results and Estimation of the Potential Impact of Individual Risk-Orientated Two-Stage First Trimester Screening", *Ultrasound in Obstetrics and Gynecology*, Vol. 25, pp. 221-227, 2005.

6. Haddow, J. E., G. E., Palomaki, G. J., Knight, J., Williams, W. A., Miller, A., Johnson, "Screening of Maternal Serum for Fetal Down's Syndrome in the First Trimester", *New England Journal Medical*, Vol. 338, pp. 955-961, 1998.

7. Hook, E. B., "Rates of Chromosome Abnormalities at Different Maternal Ages", *Obstetrics and Gynecology*, Vol. 58, pp. 282–287, 1981.

8. Cuckle, H. S., N. J., Wald, S. G., Thompson, " Estimating a Woman's Risk of Having a Pregnancy Associated with Down's Syndrome Using Her Age and Serum Alpha-fetoprotein Level", *British Journal of Obstetrics and Gynecology*, Vol. 94, pp. 387–402, 1987.

9. Merkatz, I. R., H. M., Nitowsky, J. N., Macri, W. E., Johnson, "An Association Between Low Maternal Serum Alpha-fetoprotein and Fetal Chromosome Abnormalities", *American Journal of Obstetrics and Gynecology*, Vol. 148, pp. 886–894, 1984.

10. Kuller, J. A., S. A., Laifer, "Contemporary Approaches to Prenatal Diagnosis", *American Family Physician*, Vol. 52, pp. 2277–2283, 1995.

11. Tolmie, J. L., *Emery and Rimoin's Principles and Practice of Medical Genetics*, 5th Edition, Churchill Livingstone, London, 2006.

12. Krantz, D. A., J. W., Larsen, P. D., Buchanan, "First Trimester Down Syndrome Screening: Free Beta Human Chorionic Gonadotrophins and Pregnancy Associated Plasma Protein A", *American Journal of Obstetrics and Gynecology*, Vol. 174, pp. 612-618, 1996.

13. Wald, N. J., L., George, D., Smith, "Serum Screening For Down's Syndrome Between 8 and 14 Weeks of Pregnancy", *British Journal of Obstetrics and Gynecology*, Vol. 103, pp. 407-418, 1996.

14. Palomaki, G. E., G. J., Knight, J. E., McCarthy, J. E., Haddow, J. M., Donhowe, "Maternal Serum Screening for Down Syndrome in the United States: a 1995 Survey", *American Journal of Obstetrics and Gynecology*, Vol. 176, pp. 1046–1051, 1997.

15. Saller, D. N., J. A., Canick, "Maternal Serum Screening for Down Syndrome: Clinical Aspects", *Clinical Obstetrics and Gynecology*, Vol. 39, pp.783–792, 1996.

16. Haddow, J. E., G. E., Palomaki, G. J., Knight, G. C., Cunningham, L. S., Lustig, P. A., Boyd, "Reducing the Need for Amniocentesis in Women 35 Years of Age or Older with Serum Markers for Screening", *The New England Journal of Medicine*, Vol. 330, pp. 1114–1122, 1994.

17. American College of Medical Genetics Clinical Practice Committee, "Position Statement on Multiple Marker Screening in Women 35 and Older", *American College of Medical Genetics College Newsletter*, January 1994.

18. American College of Medical Genetics Clinical Practice Committee, "Statement on Multiple Marker Screening in Pregnant Women", *American College of Medical Genetics College Newsletter*, No. 6, January 1996.

19. American College of Obstetricians and Gynecologists, "Maternal Serum Screening", *American College of Obstetricians and Gynecologists Educational Bulletin*, No. 228, 1996.

20. Benn, P. A., A., Borgida, D., Horne, S., R., Briganti, J., Rodis, "Down Syndrome and Neural Tube Defect Screening: the Value of Using Gestational Age by Ultrasonography", *American Journal of Obstetrics and Gynecology*., Vol. 176, pp. 1056–1061, 1997.

21. Chitty, L. S., "Antenatal Screening for Aneuploidy", *Current Opinion in Obstetrics and Gynecology*, Vol. 10, pp. 91–97, 1998.

22. Snijders, R. J. M., P., Noble, N., Sabire, "UK Multicentre Project on 211 Assessment of Risk of Trisomy 21 by Maternal Age and Fetal Nuchal Translucency Thickness at 10-14 Weeks Gestation", *Lancet*, Vol. 351, pp. 343-349, 1998.

23. Wald, N. J., A. K., Hackshaw, “Combining Ultrasound and Biochemistry in First Trimester Screening for Down’s Syndrome”, *Prenatal Diagnosis*, Vol. 17, pp. 821-830, 1997.

24. Spencer, K., V., Souter, N., Tul, “A Screening Programme for Trisomy 21 at 10-14 Weeks Using Fetal Nuchal Translucency, Maternal Serum Free Beta Human Chorionic Gonadotrophin and Pregnancy Associated Plasma Protein-A”, *Ultrasound in Obstetrics and Gynecology*, Vol. 13, pp.231-238, 1999.

25. Wald, N. J., C., Rodeck, K., Hackshaw, “First and Second Trimester Antenatal Screening for Down’s Syndrome: the Results of the Serum, Urine and Ultrasound Screening Study ”, *Journal of Medical Screening*, Vol. 7, pp. 1-77, 2003.

26. Smith, F., J. R., Yates, “Maternal Age Specific Rates for Chromosome Aberrations and Factors”, *Prenatal Diagnosis*, Vol. 4, pp. 5-44, 1984.

27. Merkatz, I. R., H. M., Nitowsky, J. N., Macri, W. E., Johnson, ”An Association Between Low Maternal Serum Alpha-fetoprotein and Fetal Chromosomal Abnormalities*”, American Journal of Obstetrics and Gynecology*, Vol. 148, pp. 886-894, 1984.

28. Wald, N. J., J. W., Densem, L., George, S., Muttukrishna, P. G., Knight, “Prenatal Screening for Down's Syndrome Using Inhibin-A as a Serum Marker”, *Prenatal Diagnosis*, Vol. 53, pp. 16-23, 1996.

29. Wald, N. J., C., Rodeck, A. K., Hackshaw, J., Walters, L., Chitty, A. M., Mackinson, “First and Second Trimester Antenatal Screening for Down's Syndrome: the Results of the Serum, Urine and Ultrasound Screening Study (SURUSS)”, *Journal of Medical Screening*, Vol. 104, pp. 10-56, 2003.

30. Malone, F. D., J.A., Canick, R. H., Ball, D. A., Nyberg, "First Trimester or Second Trimester Screening, or Both, for Down's Syndrome", *The New England Journal of Medicine*, 2005.

31. Brizot, M. L., R. J., Snijders, J., Butler, N. A., Bersinger, K. H., Nicolaides, "Maternal Serum hCG and Fetal Nuchal Translucency Thickness for the Prediction of Fetal Trisomies in the First Trimester of Pregnancy", *British Journal of Obstetrics and Gynecology*, Vol. 32, pp. 102-127, 1995.

32. Bromley, B., F. D., Frigoletto, B. R., Benacerraf, "Mild Fetal Lateral Cerebral Ventriculomegaly: Clinical Course and Outcome", *American Journal of Obstetrics and Gynecology*, Vol. 164, pp. 863-871, 1991.

33. Nyberg, D. A., V. L., Souter, A., El-Bastawissi, S., Young, F., Luthhardt, D. A., Luthy, "Isolated Sonographic Markers for Detection of Fetal Down Syndrome in the Second Trimester of Pregnancy", *Journal Ultrasound Medicine*, Vol. 20, pp. 1053-1060, 2001.

34. Wald, N. J., C., Rodeck, A. K., Hackshaw, J., Walters, L., Chitty, A. M., Mackinson, "First and Second Trimester Antenatal Screening for Down's Syndrome: the Results of the Serum, Urine and Ultrasound Screening Study (SURUSS)", *Journal Medicine Screen*, Vol. 10, pp. 56-66, 2003.

35. Malone, F. D., J. A., Canick, R. H., Ball, D. A., Nyberg, C. H., Comstock, R., Bukowski, "First Trimester or Second Trimester Screening, or both, for Down's Syndrome", *The New England Journal of Medicine*, Vol. 353, pp. 2001-2011, 2005.

36. Rosen, T., "Semin Perinatol", *Prenatal Diagnosis*, Vol. 29, pp.367-375, 2005.

37. Cicero, S., P., Curcio, A., Papageorghiou, "Absence of Nasal Bone in Fetuses with Trisomy 21 at 11-14 Weeks of Gestation: an Observational Study", *Lancet*, Vol. 358, pp. 1665-1673, 2001.

38. Matias, A., C., Gomes, N., Flack, “Screening for Chromosomal Abnormalities at 11-14 Weeks: the Role of Ductus Venosus Blood Flow”, *Ultrasound in Obstetrics and Gynecology*, Vol. 2, pp. 380-384, 1998.

39. Huggon, I. C., D. B., DeFigueiredo, L. D., Allan, “Tricuspid Regurgitation in the Diagnosis of Chromosomal Anomalies in the Fetus at 11-14 Weeks of Gestation”, *Heart*, Vol. 89, pp. 1071-1079, 2003.

40. Huang, K., H., Yang, I., King, M., Lyu, “Maximizing Sensitivity in Medical Diagnosis Using Biased Minimax Probability Machine", *IEEE Transactions on Biomedical Engineering*, Vol. 53, pp. 821-831, 2006.

41. Mena, L., J., Gonzalez, “Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic", *19th International FLAIRS Conference (FLAIRS- 2006)*, Melbourne Beach, Florida, May 2006.

42. Kubat, M., S., Matwin, “Addressing the Curse of Imbalanced Training Sets: One-sided Selection", *Fourteenth International Conference on Machine Learning*, pp. 179-186, Morgan Kaufmann, San Francisco, 1997.

43. Ling, C., C., Li, “Data Mining for Direct Marketing: Problems and Solutions", *Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98)*, pp. 73-79, AAAI Press, Menlo Park, CA, 1998.

44. Chawla, N., K., Bowyer, L., Hall, W., Kegelmeyer, “SMOTE: Syntethic Minority Over-Sampling Technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321-357, 2002.

45. Maloof, A. M., “Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown", *Workshop on Learning from Imbalanced Data Sets*, 2003.

46. Provost, F., “Machine Learning from Imbalanced Data Sets 101", Working Notes AAAI00 Workshop Learning from Imbalanced Data Sets, pp. 1-3, 2000.

47. Brouwer, R., “A Hybrid Neural Network with Fuzzy Rules for Categorical and Numeric Input", *International Journal of Intelligent Systems*, Vol. 19, pp. 979-1001, 2004.

48. Rogovschi, N., M., Lebbah, Y., Bennani, “Probabilistic Mixed Topological Map for Categorical and Continuous Data", *Seventh International Conference on Machine Learning and Applications*, 2008.

49. Orsenigo, C., C., Vercellis, “Predicting HIV Protease-Cleavable Peptides by Discrete Support Vector Machines", *EvoBIO*, 2007.

50. Ninomiya, T., “Clustering Observations Using Fuzzy Similarities Between Ordered Categorical Data", *Systems, Man and Cybernetics, IEEE International Conference*, 2005.

51. Lucas, P., L., Gaag, A., AbuHanna, “Bayesian Networks in Biomedicine and Health Care", *Artificial Intelligence in Medicine*, Vol. 30, pp. 201-214, 2004.

52. Gaag, L., S., Renooij, A., Feelders, A., Groote, M., Eijkemans, F., Broekmans, B., Fauser, “Aligning Bayesian Network Classifiers with Medical Contexts", Technical Report, Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands, 2008.

53. Clinical Epidemiology, *Clinical Epidemiology and Evidence-based Medicine Glossary: Clinical Study Design and Methods Terminology*, 2009, http://www.vetmed.wsu.edu/courses-jmgay/glossclinstudy.htm, accessed at May 2012.

54. Varol, F., Ö., Özer, *The Assessment of Ductus Venosus Doppler Measurements with Nukal Translucency and Serum Markers for Down Syndrome Screening in the First Trimester*, PhD Thesis, Trakya University, 2006.

55. Wapner, R., E., Thom, J. L., Simpson, E., Pergament, J., Zachary, "for the First Trimester Maternal Serum Biochemistry and Fetal Nuchal Translucency Screening (BUN) Study Group", *The New England Journal of Medicine*, Vol. 349, No. 15, 2003.

56. Alpaydin, E., *Introduction to machine learning*, Second Edition, MIT Press, Cambridge, 2010.

57. Han, J., M., Kamber, *Data mining concepts and techniques*, Academic Press, Waltham, 2001.

58. Cunningham, P., *Dimension Reduction*, University College Dublin Press, Dublin, 2007.

59. Acır, N., Ö., Özdamar, C., Güzeliş, "Automatic Classification of Auditory Brainstem Responses Using SVM-based Feature Selection Algorithm for Threshold Detection", *Engineering Applications of Artificial Intelligence* Vol. 19, pp. 209-218, 2006.

60. Duan K., J. C., Rajapakse, "Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data", *NanoBioscience IEEE Transactions*, 2005.

61. Berger, H., D., Merkl, M., Dittenbach, "Exploting Partial Decision Trees for Feature Subset Selection in e-Mail Categorization", *Proceedings of the 2006 ACM Symposium on Applied Computing*, 2006.

62. Chen, X., J. C., Jeong, "Enhanced Recursive Feature Elimination", *IEEE Sixth International Conference on Machine Learning and Applications*, 2007.

63. Guyon, I., J., Weston, S., Barnhill, V., Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, Vol. 46, Vol. 1, pp. 389–422, 2002.

64. Thang, Y., Y., Zhang, Z., Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis", *IEEE/ACM Transactions On Computational Biology and Bioinformatics*, 2007.

65. Furlanello, C., M., Serafini, S., Merler, G., Jurman, "Gene Selection and Classification by Entropy-based Recursive Feature Elimination", *Proceedings of the International Joint Conference on Neural Networks*, 2003.

66. WEKA, *WEKA 3.7: Datamining Software in Java*, 2011, http://www.cs.waikato.ac.nz/ml/weka, accessed at January 2012.

67. Quinlan, J. R., *C4.5: Programs for machine learning*, Morgan Kaufman, San Francisco, 1993.

68. Tsai, F. S., K. L., Chan, "Dimensionality Reduction Techniques for Data Exploration", *IEEE 6th International Conference on Information, Communications and Signal Processing*, Singapore, 2007.

69. Hiden, H. G., M. J., Willis, M. T., Tham, P., Turner, G. A., Montague, "Nonlinear Principal Components Analysis Using Genetic Programming", *Second International Conference On Genetic Algorithms in Engineering Systems: Innovations and Applications (GALESIA)*, 1997.

70. Lee, J. K., K. H., Kim, T. Y., Kim, W. H., Choi, "Nonlinear Principle Component Analysis Using Local Probability", *The 7th Korea Russia International Symposium on Science and Technology*, 2003.

71. Takiguchi, T., Y., Ariki, “Robust Feature Extraction Using Kernel PCA”, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

72. Wu, N., J., Zhang, “Factor Analysis Based Anomaly Detection”, *Proceedings of the 2003 IEEE Workshop on Information Assurance United States Military Academy*, West Point New York, 2003.

73. Oreški, D., P., Peharda, “Application of Factor Analysis in Course Evaluation”, *Proceedings of the ITI 2008 30th International Conerence on Information Technology Interfaces*, Cavtat, Croatia, 2008

74. IBM SPSS Software, *Predictive analytics software and solutions*, 2010, http://www-01.ibm.com/software/analytics/spss/, accessed at January 2012.

75. Lessmann, S., B., Baesens, C., Mues, S., Pietsch, “Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings", *IEEE Transactions on Software Engineering*, Vol. 34, pp. 485-496, 2008.

76. Viaene, S., R., Derrid, B., Baesens, G., Dedene, “A Comparison of State-of-the-Art Classification for Expert Automobile Insurance Claim Fraud Detection", *The Journal of Risk and Insurance*, Vol. 69, pp. 373-421, 2002.

77. Olson, D. L., D., Delen, “Advanced Data Mining Techniques”, *Springer*, 2008.

78. Marsland, S., *Machine Learning: An Algorithmic Perspective*, Chapman and Hall/CRC, London, 2009.

79. Harry Z., "The Optimality of Naive Bayes", *FLAIRS*, 2004.

80. Heckerman, D., “A Tutorial on Learning With Bayesian Networks", Technical Report, Microsoft Research Advanced Technology Division Microsoft Corporation, 1996.

81. Reiz, B., L., Csat, “Tree-Like Bayesian Network Classifiers for Surgery Survival Chance Prediction", *International Journal of Computers, Communications and Control*, Vol. 3, pp. 470-474, 2008.

82. Cheng, J., R., Greiner, J., Kelly, D., Bell, W., Liu, “Learning Bayesian Networks from Data: An Information-Theory Based Approach", *Artificial Intelligence*, Vol. 137, pp. 43-90, 2002.

83. Meloni, A., A., Ripoli, V., Positano, L., Landini, “Mutual Information Pre-conditioning Improves Structure Learning of Bayesian Networks From Medical Databases", *IEEE Trans. On Information Technology In Biomedicine*, Vol. 13, pp. 984-989, 2009.

84. Lucas, P., “Restricted Bayesian Network Structure Learning", *Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing*, pp. 217-232, Springer-Verlag, 2002.

85. Su, J., H., Zhang, C., Ling, S., Matwin, “Discriminative Parameter Learning for Bayesian Networks", *25 th International Conference on Machine Learning (ICML)*, 2008.

86. Greiner, R., and W., Zhou, “Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers", *AAAI/IAAI*, 2002.

87. Junli C., J., Licheng, “Classification Mechanism of Support Vector Machines”, *Proceedings of 5th International Conference on Signal Processing (WCCC-ICSP)*, 2000.

88. Damasevicius, R., "Optimization of SVM Parameters for Promoter Recognition in DNA Sequences", *20th EURO Mini Con-ference,Continuous Optimization and Knowledge-Based Technologies (EurOPT-2008)*, 2008.

89. Jung, T., D., Polani, "Sequential Learning with LS-SVM for Large-Scale Data Sets", *ICANN*, 2006.

90. Özkaya, A. U., *Assessing and Enhancing Machine Learning Methods in IVF Process: Predictive modeling of implantation and blastocyst development*, PhD Thesis, Boğaziçi University, 2011.

91. Bishop, C., *Pattern Recognition and Machine Learning*, Springer, Heidelberg, 2006.

92. Johansson, S., M., Jern, J., Johansson, "Interactive Quantification of Categorical Variables in Mixed Data Sets", *Proceedings of IEEE International Conference on Information Visualisation*, pp. 3-10, 2008.

93. Nukoolkit, C., H., Chen, D., Brown, "A Data Transformation Technique for Car Injury Prediction", *Proceedings of the 39th Annual ACM-SE Conference*, 2001.

94. Moturu, S., H., Liu, W., Johnson, "Healthcare Risk Modeling for Medicaid Patients", *International Conference on Healthcare Informatics*, pp. 126-133, Madeira, Portugal, January 2008.

95. Nicopoullos, J. D. M., C., Gilling, P. A., Almeida, S., Homa, L., Nice, H., Tempest, J., Ramsay, "The Role of Sperm Aneuploidy as a Predictor of the Success of Intracytoplasmic Sperm Injection?", *Human Reproduction*, Vol. 23, pp. 240-250, 2007.

96. Marble, R. P., J. C., Healy, "A Neural Network Approach to the Diagnosis of Morbidity Outcomes in Trauma Care", *Artificial Intelligence in Medicine*, Vol. 15, pp. 299-307, 1999.

97. Fawcett, T., “An Introduction to ROC Analysis", *Pattern Recognition Letters*, Vol. 27, pp. 861-874, 2006.

98. Frank, A., A., Asuncion, *UCI Machine Learning Repository*, 2010, http://archive.ics.uci.edu/ml, accessed at January 2012.