Big data analytics for early detection of breast cancer based on machine learning

Desislava Ivanova

Citation: AIP Conference Proceedings **1910**, 060016 (2017);
View online: https://doi.org/10.1063/1.5014010
View Table of Contents: http://aip.scitation.org/toc/apc/1910/1
Published by the American Institute of Physics

Big DataAnalytics for Early Detection of Breast Cancer Based on Machine Learning

Desislava Ivanova

Bulgaria, Sofia 1000, Bul. "Kliment Ohridski" 8, Technical University of Sofia,
Faculty of Applied Mathematics and Informatics,
Department of Informatics, bl. 2, office 2541

Email: d_ivanova@tu-sofia.bg

Abstract. This paper presents the concept and the modern advances in personalized medicine that rely on technology and review the existing tools for early detection of breast cancer. The breast cancer types and distribution worldwide is discussed. It is spent time to explain the importance of identifying the normality and to specify the main classes in breast cancer, benign or malignant. The main purpose of the paper is to propose a conceptual model for early detection of breast cancer based on machine learning for processing and analysis of medical big dataand further knowledge discovery for personalized treatment. The proposed conceptual model is realized by using Naive Bayes classifier. The software is written in python programming language and for the experiments the Wisconsin breast cancer database is used. Finally, the experimental results are presented and discussed.

INTRODUCTION

Most often today, a medical treatment plan doesn't have all that much to do with a patient specifically. It's identical to what doctors would hand over to essentially anyone with the same condition. That's because medicine as we know it revolves around "standards of care", the best courses of prevention or treatment for the general population. With respect to breast cancer, those standards mean self-exams and mammograms after a set age and the usual chemotherapy to treat a tumor if it is found. If the first treatment doesn't work, doctors and patients move on to the next one and the next. It's trial and error, with life on the line. Thus a more specialized approach to the patient's treatment is required [1].

The right manner and the solution for those cases is the genomic or personalized medicine. It is the use of information from genomes and their derivatives (RNA, proteins, and metabolites) to guide medical decision making that is a key component of personalized medicine, which is a rapidly advancing field of healthcare that is informed by each person's unique clinical, genetic, genomic, and environmental information. As medicine begins to embrace genomic tools that enable more precise prediction and treatment disease, which include whole genome interrogation of sequence variation, transcription, proteins, and metabolites, the fundamentals of genomic and personalized medicine will require the development, standardization, and integration of several important tools into health systems and clinical workflows [2].

Every person has a unique variation of the human genome. Although most of the variation between individuals has no effect on health, an individual's health stems from genetic variation with behaviors and influences from the environment. Modern advances in personalized medicine rely on technology that confirms a patient's fundamental biology, DNA, RNA, or protein, which ultimately leads to confirming disease [3, 4].

The concept of personalized medicine can be applied to new and transformative approaches to healthcare. It can also be used to predict a person's risk for a particular disease, based on one or even several genes. All these activities realizing the concept of personalized medicine generate a large amount of data (big medical datasets) that need to be processed and analyzed. The solution to processed and analyze the big data for early detection of disease is the usage of new scientific paradigm "Data-Intensive Scientific Discovery (DISD).

Proceedings of the 43rd International Conference Applications of Mathematics in Engineering and Economics
AIP Conf. Proc. 1910, 060016-1–060016-8; https://doi.org/10.1063/1.5014010
Published by AIP Publishing. 978-0-7354-1602-4/$30.00

The fourth paradigm triggered a revolution in scientific research and innovations, involving accumulation of data, dataanalysis and data-intensive decision making. It imposes a set of challenges towards the processing technologies as accumulation and storage of huge amounts of data, analysis and visualization, high performance processing resources, parallel and distributed processing [5].

This paper will proposed aconceptual model for early detection of breast cancer. In order to solve all challenges connecting to the proposed approach, machine learning techniques will be applied to process and analyze the medical big dataand to extract the knowledge for furtherpersonalized treatment.

BREAST CANCER: TYPES, STATICTICS AND TOOLS FOR EARLY DETECTION

Breast Cancer Types and Stats

Breast cancer starts when cells in the breast begin to grow out of control. These cells usually form a tumor that can often be seen on an x-ray or felt as a lump. The tumor is malignant (cancerous) if the cells can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Breast cancer occurs almost entirely in women, but men can get it, too. Breast cancers can start from different parts of the breast. Most breast cancers begin in the ducts that carry milk to the nipple (ductal cancers). Some start in the glands that make breast milk (lobular cancers) [6].

In generalat the first stage, the oncology experts try to describe whether the cancer has spread beyond the place it started. According this characteristics, there are two main categoriesof breast cancers: *in situ breast cancers* which have not spread and *invasive cancers* which have invaded into the surrounding breast tissue.There are a lot of special types of invasive breast carcinomalike: *adenoid cystic carcinoma, low-grade adenosquamous carcinoma, medullary carcinoma, mucinous carcinoma, papillary carcinoma, tubular carcinoma.* There are less common types of breast cancer like *inflammatory breast cancer, Paget disease of the nipple, Phyllodes tumorand angiosarcoma [7].*

It is really important to understand that most breast lumps are not cancer, they are benign. Benign breast tumors are abnormal growths, but they do not spread outside of the breast and they are not life threatening. But some benign breast lumps can increase a woman's risk of getting breast cancer. Any breast lump or change needs to be checked by a health care provider to determine whether it is benign or cancer, and whether it might impact the future cancer risk. Breast cancer can spread through the lymph system.

The lymph system includes lymph nodes, lymph vessels and lymph fluid found throughout the body. Lymph nodes are small, bean-shaped collections of immune system cells that are connected by lymph (or lymphatic) vessels. Lymph vessels are like small veins, except that they carry a clear fluid called lymph (instead of blood) away from the breast. Lymph contains tissue fluid and waste products, as well as immune system cells. Breast cancer cells can enter lymph vessels and begin to grow in lymph nodes. If cancer cells have spread to your lymph nodes, there is a higher chance that the cells could have spread (metastasized) toother sites in your body. The more lymph nodes with breast cancer cells, the more likely it is that the cancer may be found in other organs as well. Because of this, finding cancer in one or more lymph nodes often affects the treatment plan. Usually, surgery to remove one or more lymph nodes will be needed to know whether the cancer has spread there.Still, not all women with cancer cells in their lymph nodes develop metastases, and some women can have no cancer cells in their lymph nodes and later develop metastases [8].

The early detection is really important for the human life especially for the invasive breast carcinoma types. According toWorld Cancer Research Fund International (WCRFI), the statistics shows that breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012. It is the second most common cancer overall. This represents about 12% of all new cancer cases and 25% of all cancers in women.

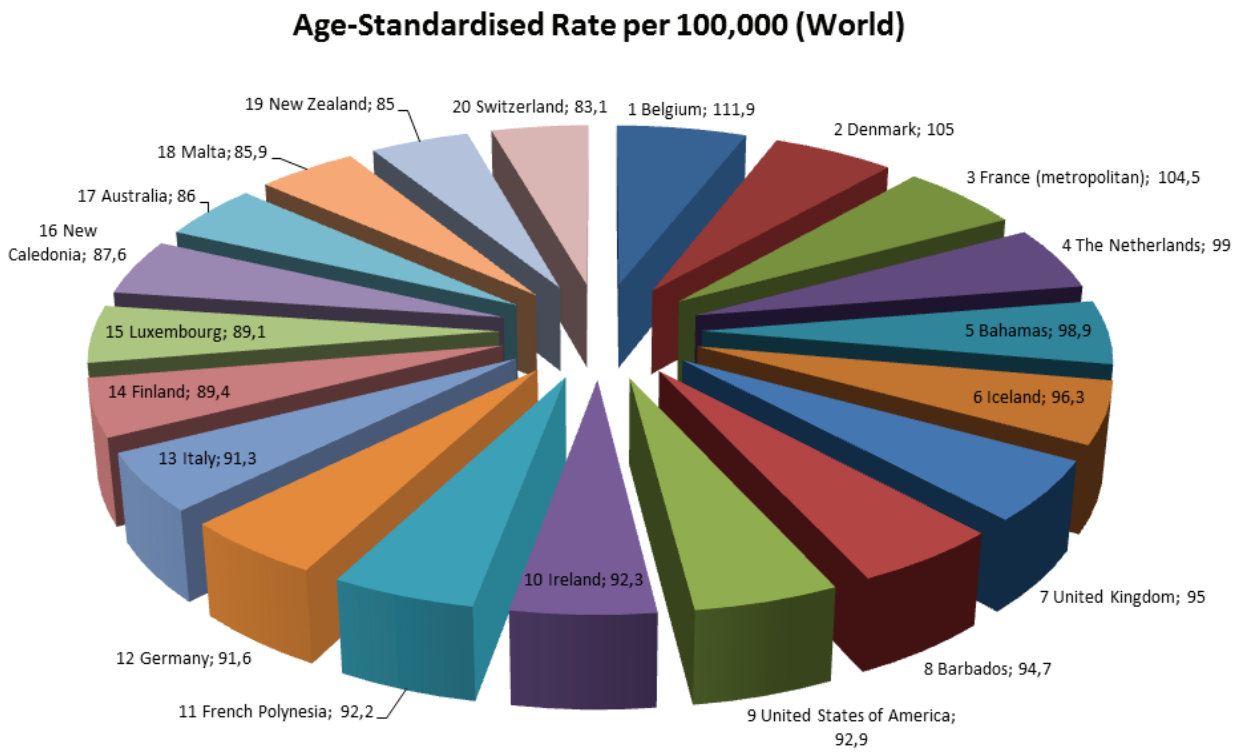

FIGURE 1. Preventing Breast Cancer Stats according WCRFI

The countries with the top 20 highest incidence of breast cancer according WCRFIare presented on Fig. 1. The data show that Belgium had the highest rate of breast cancer followed by Denmark and France.Slightly more cases of breast cancer 53% were diagnosed in less developed countriesand the highest incidence of breast cancer was in Northern Americaand Oceaniaand respectively the lowest incidence in Asiaand Africa [9].

According to Patient Association of Oncological Diseases in Bulgaria, the new breast cancer cases entered in the National Cancer Registry are between 3600 and 3800.In 1992, the diagnosed Bulgarian breast cancer patients are 26,000 and in 2012, they are over 34,000. It is alarming that breast cancer patients in Bulgariaare increasing - 34,000 a decade ago up to 44,425 nowadays. 20% of these or nearly 8500 people are at an advanced stage with metastases in bone, brain, and other. Late diagnosis and inadequate treatment in most cases is fatal.

Tools for Early Detection of Breast Cancer

Tests and exams used to find a disease, like cancer, in people who do not have any symptoms are called screening tests. Screening exams, such as mammograms, find cancers before they start to cause symptoms. This is called early detection. Cancers that are found early, when they're small and haven't spreadare easier to treat and have better outcomes. Breast cancers that are found because they can be felt tend to be larger and are more likely to have already spread outside the breast. But screening exams can often find breast cancers when they are small and still confined to the breast. The size of a breast cancer and how far it has spread are some of the most important factors in predicting the outlook (prognosis) of a woman with this disease. Screening tests are used to find breast cancer before it causes any warning signs or symptoms. Screening tests can find breast cancer early, when the chances of survival are highest. Regular screening tests (along with follow-up tests and treatment if diagnosed) reduce your chance of dying from breast cancer. A clinical breast exam is a physical exam done by a health care provider. It is often done during your regular medical check-up. The provider will visually check your breasts while you are sitting up and physically examine your breasts while you are lying down.

Having the ability to look at a patient on an individual basis will allow for a more accurate diagnosis and specific treatment plan. Genotyping is the process of obtaining an individual's DNA sequence by using biological assays. By having a detailed account of it, the genome can then be compared toa reference genome, like that of the Human

Genome Project, toaccess the existing genetic variations that can account for possible diseases. Having this information from individuals can then be applied to effectively treat them. An individual's genetic make-up also plays a large role in how well they respond toa certain treatment, and therefore, knowing their genetic content can change the type of treatment they receive [10].

In addition to specific treatment, personalized medicine can greatly aid the advancements of preventive care. For instance, many women are already being genotyped for certain mutations in the BRCA1 and BRCA2 gene if they are predisposed because of a family history of breast cancer or ovarian cancer. As more causes of diseases are mapped out according to mutations that exist within a genome, the easier they can be identified in an individual. Measures can then be taken to prevent a disease from developing. Even if mutations were found within a genome, having the details of their DNA can reduce the impact or delay the onset of certain diseases. Having the genetic content of an individual will allow better guided decisions in determining the source of the disease and thus treating it or preventing its progression. This is extremely useful for diseases like cancersand especially for breast cancer that are thought to be linked to certain mutations in the DNA.

BRCA1 and BRCA2 are human genes that produce tumor suppressor proteins. These proteins help repair damaged DNAand, therefore, play a role in ensuring the stability of the cell's genetic material. When either of these genes is mutated, or altered, such that its protein product either is not made or does not function correctly, DNA damage may not be repaired properly. As a result, cells are more likely to develop additional genetic alterations that can lead to cancer. Specific inherited mutations in BRCA1 and BRCA2 increase the risk of female breast and ovarian cancers, and they have been associated with increased risks of several additional types of cancer. Together, BRCA1 and BRCA2 mutations account for about 20 to 25 percent of hereditary breast cancers and about 5 to 10 percent of all breast cancers. In addition, mutations in BRCA1 and BRCA2 account for around 15 percent of ovarian cancers overall. Breast and ovarian cancers associated with BRCA1 and BRCA2 mutations tend to develop at younger ages than their nonhereditary counterparts [11].

A harmful BRCA1 or BRCA2 mutation can be inherited from a person's mother or father. Each child of a parent who carries a mutation in one of these genes has a 50 percent chance (or 1 chance in 2) of inheriting the mutation. The effects of mutations in BRCA1 and BRCA2 are seen even when a person's second copy of the gene is normal. Harmful mutations in BRCA1 and BRCA2 increase the risk of several cancers in addition to breast and ovarian cancer. BRCA1 mutations may increase a woman's risk of developing fallopian tube cancer and peritoneal cancer. Men with BRCA2 mutations, and toa lesser extent BRCA1 mutations, are alsoat increased risk of breast cancer. Men with harmful BRCA1 or BRCA2 mutations have a higher risk of prostate cancer [12].

In order for physicians to know if a mutation is connected toa certain disease, researchers often doa study called a "genome-wide association study" (GWAS). A GWAS study looks at one disease, and then sequence the genome of many patients with that particular disease to look for shared mutations in the genome. Mutations that are determined to be related toa disease by a GWAS study can then be used to diagnose that disease in future patients, by looking at their genome sequence to find that same mutation. The first GWAS, conducted in 2005, studied patients with age-related macular degeneration (ARMD). It found two different mutations, each containing only a variation in only one nucleotide (called single nucleotide polymorphisms, or SNPs), which were associated with ARMD. GWAS studies like this have been very successful in identifying common genetic variations associated with diseases [13].

Over recent decades cancer research has discovered a great deal about the genetic variety of types of cancer that appear the same in traditional pathology. There has also been increasing awareness of tumor heterogeneity, or genetic diversity within a single tumor. Among other prospects, these discoveries raise the possibility of finding that drugs that have not given good results applied toa general population of cases may yet be successful for a proportion of cases with particular genetic profiles.

Cancer Genomics is the toolof genomics and personalized medicine to cancer research and treatment. High-throughput sequencing methods are used to characterize genes associated with cancer to better understand disease pathology and improve drug development. It is one of the most promising branches of genomics, particularly because of its implications in drug therapy.

Another aspect is pharmacogenomics, which uses an individual's genome to provide a more informed and tailored drug prescription. Often, drugs are prescribed with the idea that it will work relatively the same for everyone, but in the application of drugs, there are a number of factors that must be considered. The detailed account of genetic information from the individual will help prevent adverse events, allow for appropriate dosages, and create maximum efficacy with drug prescriptions. The pharmacogenomic process for discovery of genetic variants that predict adverse events toa specific drug has been termed toxgnostics.

With respect toanalyzing RNA-seq data, differential gene expression analysis, alternative splicing analysisand integration and visualization of multiple data types, there are different software tools like NOISeq R package, RNAseqViewer, UCSC browser, Integrative Genomics Viewer (IGV), Genome Mapsor Savant but still they are not fully capable to provide the information about gene mutations and cancer detection.

BIG DATAANALYTICS FOR EARLY DETECTION OF BREAST CANCER BASED ON MACHINE LEARNING

In this paper, the conceptual model for early detection of breast cancer is proposed. Generally, the early detection of breast cancer involves various stages, as shown in Fig. It should be noted the goal of this research is to provide expert oncologist a second opinion and help them to increase the early detection accuracy, reduce biopsy rate and save the time and effort.

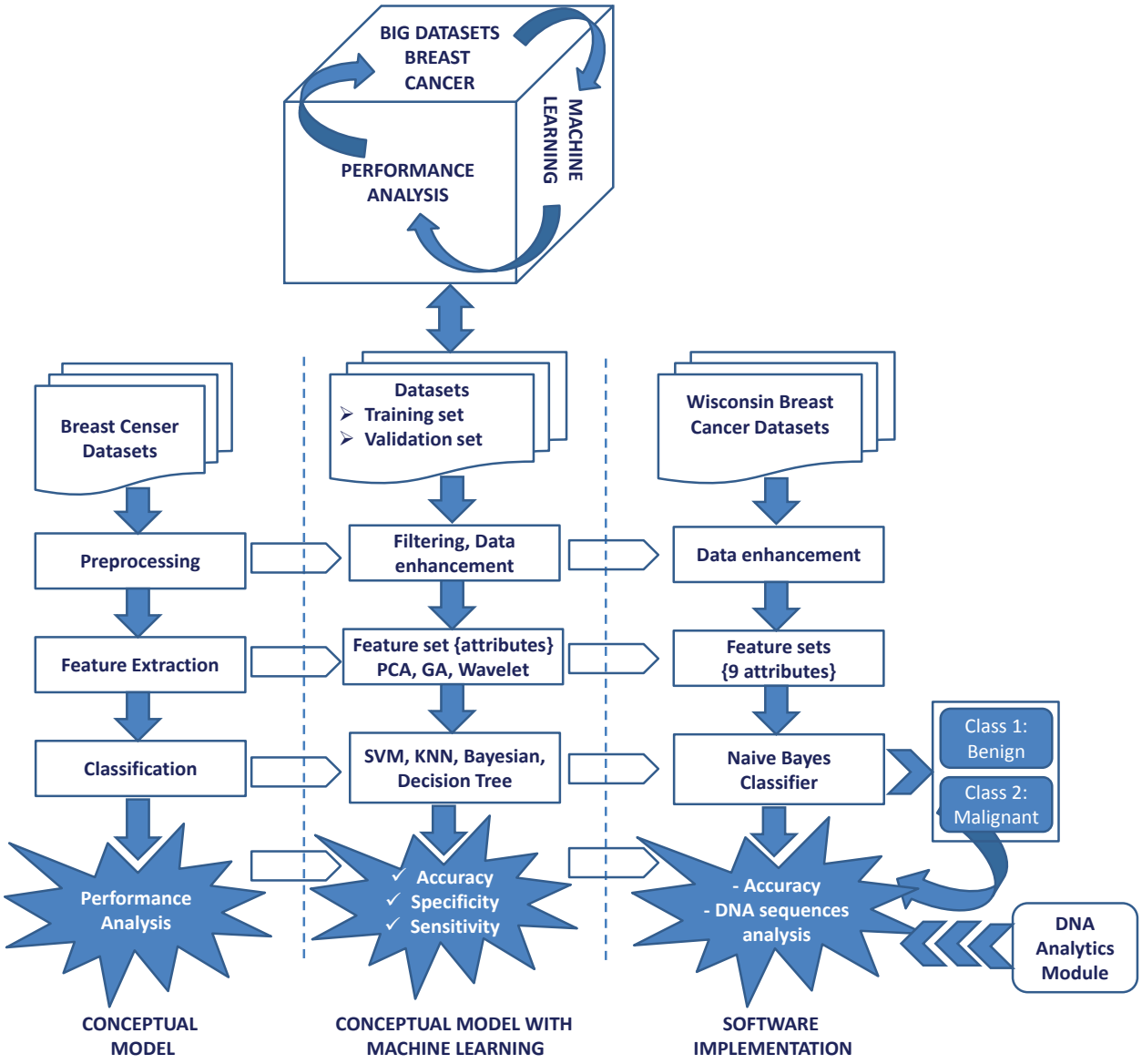

FIGURE 2. Conceptual Model for Early Detection of Breast Cancer based on Machine Learning

The basic steps of the proposed conceptual model are *pre-processing, feature extraction, classification and performance analysis.* The first step of the conceptual model takes intoaccount data cleaning (filtering) and data enhancements for processing. The second step is feature extraction. It is connecting to the process of extracting certain characteristic attributes and generating a set of meaningful descriptors from a medical datasets. It is used to find a feature set that can accurately distinguish the sample like benign or malignant.Furthermore in this step, the feature reduction by removing irrelevant or redundant data is done that has an immediate effect on application by accelerating the classification algorithm. The next step is classification, it is used to classify the data in two classes benign and malignant based on the selected features. There are different classification methods based on machine learning that can be used in order to implement the classifiers like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree and etc. Finally, the performance parameters like accuracy, sensitivity and specificity are given.

In this paper, the proposed conceptual model is realized by using Naive Bayes classifier. The software is written in python programming language and for the experiments the Wisconsin breast cancer database is used.

EXPERIMENTS AND RESULT ANALYSIS

This breast cancer databases was obtained from the University of Wisconsin Hospitals where the samples arrive periodically [14]. The database therefore reflects this chronological grouping of the data. In this database the following attributesareexisted Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. The values of the attributes are between 1 and 10. Each instance of the dataset has one of two possible classes: benign indexed with 2or malignant index with 4. The class distribution is for Benign: 458 (65.5%) and for Malignant: 241 (34.5%). The example view of the dataset is shown on the figure below, where every new case in database is represented with one row with the associating value of the attributes.

ATTRIBUTE	DOMAIN
Sample code number	id number
Clump Thickness	1 - 10
Uniformity of Cell Size	1 - 10
Uniformity of Cell Shape	1 - 10
Marginal Adhesion	1 - 10
Single Epithelial Cell Size	1 - 10
Bare Nuclei	1 - 10
Bland Chromatin	1 - 10
Normal Nucleoli	1 - 10
Mitoses	1 - 10
Class:	2 for benign 4 for malignant

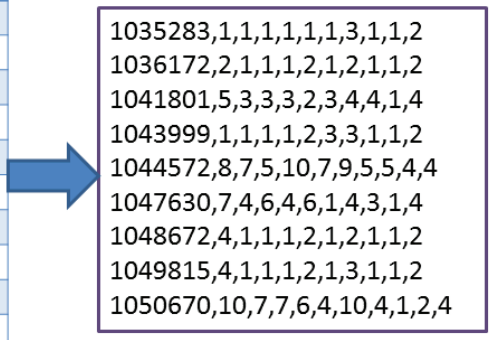

FIGURE 3. Breast cancer Wisconsin Data

The software that implements the proposed approach is written in python programming language and used the naive Bayes classifier. In machine learning, naive Bayes classifiers are a family of probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Bayes theorem provides a way of calculating the posterior probability from class prior probability, predictor prior probability and likelihood where posterior probability is the posterior probability of class (target) given predictor (attribute) and the likelihood is the probability of predictor given class.

In the software compounds of some functions, each function has a single task. The first function opens the text file with the database and reads the data, actually every single attribute from the database. The second one through the lambda expression travels over the database and verifies the truth of existing attributes in the database. In case there is missing, the program gives the Boolean value false. Then the function returns the length of single set in the

database, divided on the events, like the rule of Bayes probability. The third function takes the values of the attributes and replaces them in 2 arrays, called benignand malignant. The next function gives the prediction. At last the important function is the main function. The main function reads the database and called all results from other functions, it evaluates precision of the model then increases "if-statement". There increments the success, through the probability. At last the main function prints the resultof the naive Bayes classifiers accuracy. The value of the achieved naive Bayes classifiers accuracy is 0.978597610089. The formula that calculates the accuracy is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where: TP (true positive), FN (false negative), FP (false positive), TN (true negative).

TABLE 1. Naive Bayes Classifier Performance Results

Number of Attributes	*Features sets*	*Accuracy*
9	{Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses}	97,86 %

In addition, the software offers a module for analyzing theDNA sequences. Because for this research, it is important to know more about the breast cancer, the sequences of BRCA1 and BRCA2 genes are downloaded from NCBI in FASTA format and saved in text format. Both FASTA files areanalyzed.

In software there is afunction to read the file and to name it on "gene". Then because of the calculation the number of nucleotides, it is equalized every single nucleotide to 0. Then trough the method read line it is skipped the first line of information it is not need, because this line hasn't nucleotides. Then it increments every nucleotide and calculate the content through division on the complete content of nucleotides in this DNA sequence. The software gives three results: *the number of nucleotides in the sequence, the Adenine – Thymine content and the Guanine – Cytosine content.* The results from the analyses of both BRCA1 and BRCA2 genes are showed on the figure below.

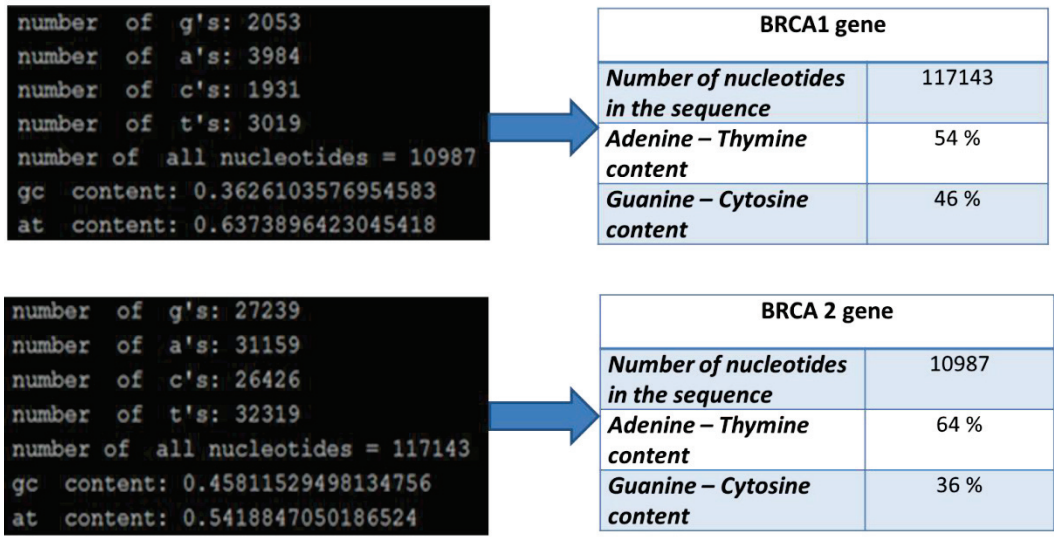

FIGURE 4. Analysis of BRCA1 and BRCA2 Genes

Those results can be further checked it in the BLAST program, local database search tool. There the program compares both sequences and when it has missing or incompatibilities nucleotides, the program put gabs.

CONCLUSION AND FUTURE WORK

A lot of research teams' worldwide applied different techniques to process medical data with respect to identifying breast cancer as many of the used dataset are hardly readable with different structures taking in account different attributes that makes the detection of normality really heavy and challenging task. It is really important to be categorized the breast cancer anomalies and to be created a relevant datasets. That is possible with the joint effort of IT and the oncologist experts. The most urgent future task is to be crated the benchmark breast cancer datasets. That will support the performance evaluation and comparison of the different algorithms results.

ACKNOWLEDGMENTS

The research is supported by the Project "Intelligent Method for Adaptive In-silico Knowledge Discovery and Decision Making Based on Analysis of Big Data Streams for Scientific Research" funded by the Bulgarian National Science Foundation, Bulgarian Ministry of Education and Science, Competition for financial support of fundamental research(2016) under the thematic priority: Technical Science, contract № ДН07/24 - 15.12.2016.

REFERENCES

1. M. Song, K.M. Lee, D. Kang, Breast cancer prevention based on gene-environment interaction. Mol. Carcinog. 2011;50:280–290. doi: 10.1002/mc.20639.
2. T. Stricker, D.V. Catenacci, S.Y. Siewert, Molecular profiling of cancer - the future of personalized cancer medicine: a primer on cancer biology and the tools necessary to bring molecular testing to the clinic. Semin. Oncol. 2011;38:173–185. doi: 10.1053/j.seminoncol.2011.01.013.
3. American Cancer Society, Personalized medicine: redefining cancer and its treatment, 2015 Report - Part 1, Personalized cancer care: Where it stands today, 2015 Report - Part 2.
4. The Royal Society, Recent developments in personalized medicine, Royal Society's report Personalized Medicine: Hopes and Realities, 2016.
5. V. M. Schönberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, 2014, www.amazon.com,eBook.
6. C. Sang-Hoon, J. Jeon, S. Kim, Personalized medicine in breast cancer: Asystematic review, 2012 Sep; 15(3): 265–272, doi: 10.4048/jbc.2012.15.3.265.
7. http://www.breastcancer.org/symptoms/types
8. U. Jayasinghe,N. Pathmanathan, E. Elder,J. Boyages, Prognostic value of the lymph node ratio for lymph-node-positive breast cancer- is it just a denominator problem?, Springerplus, 2015; 4: 121, PMCID: PMC4366431, doi: 10.1186/s40064-015-0865-2.
9. World Cancer Research Fund International: http://www.wcrf.org
10. Human Genome Project, NHGRI's Division of Intramural Research, https://www.genome.gov
11. N. Petrucelli, M.B. Daly, T. Pal, BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer, Copyright © 2017, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved, PMID: 20301425.
12. Canadian Cancer Society, BRCA gene mutations, 2017, http://www.cancer.ca
13. P. M. Visscher, M.A. Brown,M.I. McCarthy, J.Yang, Five Years of GWAS Discovery, 2012 Jan 13; 90(1): 7–24, doi: 10.1016/j.ajhg.2011.11.029.
14. William H. Wolberg, Breast Cancer Database, University of Wisconsin Hospitals, Madison, Wisconsin, USA.