INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2020.14.74 Volume 14, 2020

Platform for Big Biomedical Data Streams Management and Analytics

Veska Gancheva
Technical University of Sofia
Sofia 1000, 8 Kliment Ohridski boul.
Bulgaria

Received: May 1, 2020. Revised: July 24, 2020. 2nd Revised: September 15, 2020.
Accepted: September 18, 2020. Published: September 21, 2020.

Abstract—Major challenge in the analysis of clinical data and knowledge discovery is to suggest an integrated, advanced and efficient tools, methods and technologies for access and processing of progressively increasing amounts of data in multiple formats. The paper presents a platform for multidimensional large-scale biomedical data management and analytics, which covers all phases of data discovery, data integration, data preprocessing, data storage, data analytics and visualization. The goal is to suggest an intelligent solution as integrated, scalable workflow development environment consisting of a suite of software tools to automate the computational process in conducting scientific experiments.

Keywords—Big biomedical data analytics, knowledge discovery from data, machine learning, precision medicine.

I. Introduction

HUHE amounts of data have been generated as a result of the computer simulations. This yielded an intensive development of methods and technologies for big data processing and changes in the scientific research paradigms, such as data-intensive science. The change supposes a new research performing and knowledge discovery on the basis of data analysis. Significant correlations are sought during this process and innovative methods for knowledge discovery are applied. The methods and technologies of the new paradigm give the opportunity to process raw data in their integrity, and knowledge discovery is based on „data-intensive decision making“ [1].

The term Big Data is used to describe the rapid increase in the volume, variety and velocity of information available in almost every aspect of our lives, including medical research [2]. The scientists now have the capacity to rapidly generate, store and analyze data. The term Big data has expanded and now refers not to just large data volume, but to increasing ability to analyze and interpret those data. New methods aimed to improving data collection, storage, cleaning, processing and interpretation continue to be developed.

Big data analytics is the process of examining large data containing heterogeneous data sets. Big data analytics aims to uncover hidden patterns, unknown correlations, complex trends, imbalanced datasets, customer/clinical center preferences, as well as other useful features [3].

The accumulation and storage of huge amounts of data, generally of low information density, is becoming a major source of knowledge. The new paradigm for research is Data-Intensive Science [4] after three paradigms: 1) empirical, 2) theoretical and 3) computational. The radical change in the fourth paradigm implies a new way of conducting experiments and discovering knowledge. Instead of planning the experiment and then analyzing the data, in the new paradigm, huge amounts of data are subjected to an analysis that looks for hidden models, meaningful correlations and cause-and-effect relationships in the data, and applies innovative intelligent methods to discover new knowledge. The methods and technologies of the new paradigm enable the processing of raw data in their entirety, with the discovery of new knowledge based on “data-intensive decision making”.

The Data-Intensive Scientific Discovery (DISD) research paradigm [5] has revolutionized research and innovation, suggesting the following phases: (1) data accumulation, (2) "cleanup", data integration and transformation, (3) data analysis, and (4) data-intensive decision making. The fourth paradigm poses a number of challenges to computing technology that can be summarized as follows: accumulating and storing huge amounts of data, searching, sharing, analyzing and visualization, the need for high-performance computing products, parallel and distributed processing, parallel input/output processing in the memory. In the process of data analysis main problem is the scalability. For large data key challenge is also the assembly processing.

The new challenge is how to reveal the underlying mechanisms of biological systems by understanding big data. Today, life sciences need more efficient, computable, quantitative, accurate and precise approaches to deal with big

This work was supported by the Bulgarian National Science Fund, Ministry of Education and Science, under Grant KP-06-N37/24.

V. Gancheva is with the Programming and Computer Technologies Department, Technical University of Sofia, Bulgaria, phone: +35929652192; e-mail: vgan@ tu-sofia.bg.

data. Hypothesis-driven research is a key to knowledge discovery from big biological data.

The amount of data is so large that traditional data analysis platforms and methods can no longer meet the need to quickly perform data analysis tasks and discover knowledge in the life sciences. The big challenge is processing the combination of big data that is collected in real time and data that is already accumulated. As a result, both biologists and medical scientists and computer scientists face the challenge of getting an in-depth look at the deepest features of big biomedical data. This in turn requires enormous computing resources. Therefore, highly efficient computing platforms are needed, as well as new technologies, scalable methods and algorithms for data analysis and the presentation of open knowledge in the field of artificial intelligence such as machine learning, models based on social behavior, meta-heuristic, neural networks, topological analysis, etc. as well as models for 3D stereoscopic visualization for better understanding of knowledge discovered.

The problem of adapting the personal treatment to a patient is extremely complex and the medical doctor should examine and analyze large amounts of various data of different type. Big data technologies can be very helpful in helping medical doctors with data analysis. A comprehensive system for precision medicine, which covers all phases of data discovery, data integration, data preprocessing, building models, data storage, data analysis and visualization can be very useful to scientists in support of precision medicine. Major challenge in the analysis of clinical data is to propose an integrated and modern access to the progressively increasing amounts of data in multiple formats, and efficient approaches for their management and processing.

From the above, it can be concluded that the problem of adapting the personal treatment to a patient is extremely complex and the medical doctor should examine and analyze large amounts of various data of different type. Big data technologies can be very helpful in helping doctors with data analysis. A comprehensive platform, which covers all phases of data discovery, data integration, data preprocessing, building models, data storage, data analysis and visualization can be very useful to scientists in support of precision medicine.

The purpose of this paper is to present a platform for multidimensional large-scale biomedical data analytics and knowledge discovery. The goal is to suggest an intelligent solution as integrated, scalable workflow development environment consisting of a suite of software tools to automate the computational process in conducting scientific experiments.

The paper is structured as follows. Related work is presented in Section II. Section III describes a platform for biomedical data analytics and knowledge discovery. Section IV is focused on breast cancer classification based on machine learning algorithms. The experimental results and analyses are explained in section V.

II. Related Work

Human genetic diversity, rare and common mutations associated with sensitivity of human disease and genetic diversity are important for precision medicine [6]. Precision medicine proposes appropriate and optimal disease diagnostics, medical decision, treatment and therapy, being tailored to the individual characteristics of each patient, especially based on individual patient's genetic analysis [7].

Exploiting new tools to extract meaning from large volume information has the potential to drive real change in clinical practice, from personalized therapy and intelligent drug design to population screening and electronic health record mining. Main challenges include the need for standardization of data content, format, and clinical definitions, a heightened need for collaborative networks with sharing of both data and expertise. Data mining approach for information extraction from biomedical domain based Support Vector Machine (SVM) by adopting the use of text mining framework is developed [8].

The development of new tools and technologies for Big Data analysis and visualization has led to the rapid growth of precision medicine [9]. In the era of big data and with the development of electronic healthcare records, large and comprehensive databases of genomic, transcriptomics, proteomics or metabolomics variables, as well as traditional clinical patients' characteristics and treatment records have emerged at an increasingly rapid pace [10]. However, such data are often very heterogeneous, high dimensional, noisy and poorly interpretable in the context of their direct usage in a clinical environment such as precise diagnostic and disease outcome prediction [11]. In addition, the data quality is often limited by small sample size, imperfect technology, difference in clinical trials, diversity of patient cohorts and health care system model.

Scientists have developed new strategies for early prognosis of cancer treatment outcomes [12]. With the emergence of new technologies in the medical field, large amounts of biomedical data are collected and are available for the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for medical doctors. Artificial intelligence techniques are popular for medical researchers and allow discovering and identifying models of complex datasets and relationships between them, leading to effectively predict future outcomes of a disease type.

The development of some new biomedical technologies generating biomedical data like genomics, medical generation - sequencers that generate genetic data and imaging tools - CT imaging, nuclear magnetic resonance imaging (MRI), multilayer microscopic imaging in cell analysis, and more lead to a large amount of heterogeneous data accumulation. During the last decade, we have witnessed an explosion in the amount of the available bioinformatics data, due to the rapid progress of high-throughput sequencing projects. An unprecedented amount of data is generated daily containing clinical reports, genomic sequences, gene expression profiles, biomedical

literature reports, medical imaging and sensor data. For example, the European Institute of Bioinformatics maintains approximately 273 petabytes of raw storage data - genes, proteins and small molecule data [13].

With this exponential growth of biological data, applications for appropriate analysis are being developed and studied for the analysis of multiple biological data, such as sequencing, genome assembly, single nucleotide polymorphism detection, and genome-wide association studies. Many of these applications have several common features: 1) a huge amount of data that is generated in sequencing centers, such as Illumina being able to generate over 1.8 terabytes per week ; 2) an extremely long processing time, for example, the SOAPdenovo2 [14] genome-assembly tool takes several days to spend hundreds of GB of memory to complete the genome-building of one person; and 3) application dependency - to obtain the end result from which useful knowledge can be discovered, different processing steps need to be completed, resulting in significant operational costs for data transmission.

In the life-science era of Omics, data are presented in many forms that represent information at different levels of biological systems, including genome, transcriptome, epigenome, proteome, metabolome, molecular imaging, molecular pathways, and clinical / medical data records [15]. Data is large and their volume is already well above petabytes (PB), even exabytes (EB). No one doubts that biomedical data will create enormous amounts of value and lead to the extraction of valuable knowledge if scientists overcome many challenges, such as how to deal with the complexity and integration of data from many diverse resources. The tools and techniques for analyzing big biological data make it possible to translate the vast amount of information into a better understanding of the basic biomedical mechanisms that can be applied further to personalized medicine.

For the successful diagnosis, monitoring and treatment of diseases, medicine and technology combine to work together to develop state-of-the-art medical platforms that enable the creation and management of workflows for image processing, 3D visualization of the patient's anatomy containing high quality details, organizing, analyzing, and sharing medical images and academic research. Medical platforms like healthcare application framework for AI powered imaging and genomics NVIDIA Clara [16] and deep learning framework Caffe [17] are systems that use multiple tools that are appropriately combined and managed by the user to perform a complex process of organizing and automating data analysis.

III. Biomedical Data Analytics and Knowledge Discovery Platform

A. Requirements Design

The work presented in this paper is a part of a project that offers a scientific platform for intelligent management and analysis of big data streams supporting biomedical scientific research and precision medicine. The major advantage is the automatic generation of hypotheses and options for decisions, as well as verification and validation utilizing biomedical data set and expertise of scientists. The goal is to create an integrated open technology platform for intelligent solutions for multidimensional large-scale biomedical data analysis based on innovative and effective methods, algorithms and tools for biomedical data discovery, integration, storage and analysis, knowledge discovery and decision making for the needs of medicine. As a result, the platform is expected to offer the scientists an integrated, scalable workflow development environment consisting of a suite of software tools to automate the computational process in conducting scientific experiments and provide:

1) an easy-to-use environment;

2) interactive tools for executing workflows and visualizing real-time results through virtual reality;

3) sharing and reusing workflows;

4) ability to track workflow execution results and workflow creation steps;

5) higher efficiency and speed.

Big data infrastructure is a framework used to store, process and analyze big data. Big data analysis involves the collection, management, and analysis of massive, diverse datasets to uncover hidden patterns, correlations, causation, and other intuitions in big data infrastructure. Because of its effectiveness, big data analysis is widely used in various research fields, including biology and medicine.

Big data technology is typically associated with three perspectives on technical innovation and super-large datasets: automated parallel computing, data management schemas, and data mining. Technologically, the infrastructure for large data sets includes:

1) Scalable storage that is used to collect, management, and analyze massive datasets.

2) A computer platform configured specifically for scalable analysis that consists of multiple (usually multi-core) processing units.

3) A data management environment whose configuration can range from a traditional database management system scaled to massive concurrency to databases configured with alternative distributed systems to new graph-based or NoSQL data management schemes.

4) Scalable application development and integration framework that includes programming models, development tools, application execution, and system configuration and management capabilities.

5) Scalable analytics methods (including data mining models) to improve the ability to design and build analytical and forecast models.

6) Scientific workflows and management tools, consistent with storage, integration and analysis infrastructure. Finding and retrieving data from a variety of sources is a challenging task in itself.

Creating a common structure to store all the huge information for analysis is another challenge. The changing requirements for the analysis of big data sets have created a

more modern paradigm for data virtualization, driven by the search for easier access and federation of data - Logical Data Warehouse (LDW). An architectural layer has been added to look at the data without having to move and transform beforehand, combining the strengths of traditional repositories with alternative data management and access strategies. One of the latest approaches to analyzing big data is Data Lakes, where huge amounts of data are stored in raw format as needed. Data Lakes uses flat storage architecture, unlike Data Warehouse, which stores data in a database or hierarchical file system.

A. Biomedical Data Analytics and Knowledge Discovery Architecture

Biomedical data analytics and knowledge discovery architecture for multidimensional big biomedical data analysis based on effective methods, algorithms and tools for data integration, data preprocessing, data storage and data analysis and visualization, knowledge discovery and decision making for the purpose of precision medicine is presented in Fig. 1.

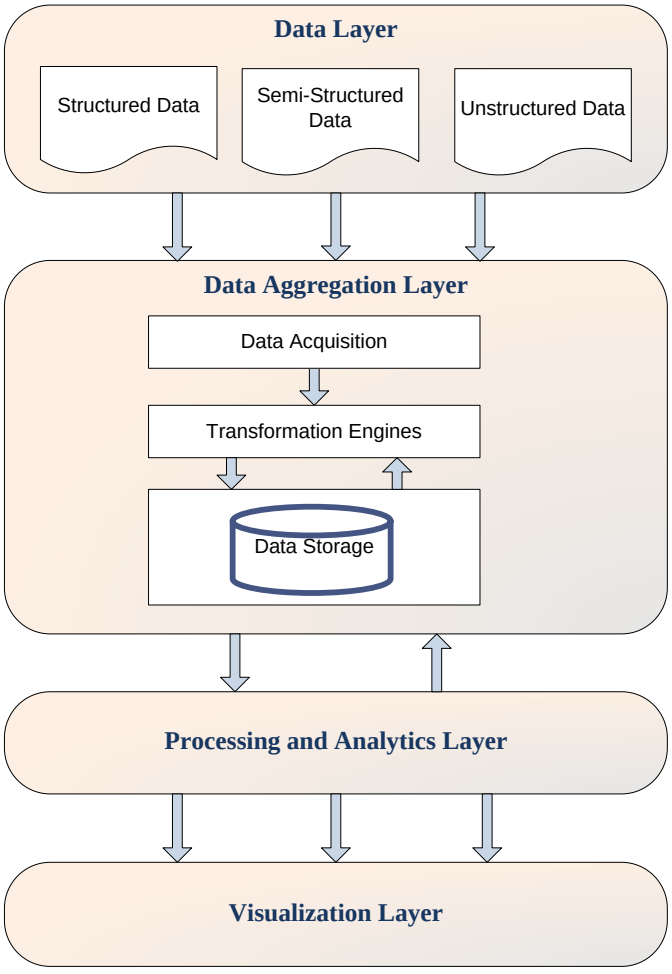

Fig. 1 Biomedical data analytics and knowledge discovery architecture

The system architecture follows the processing pipeline for discovering useful knowledge from a collection of data and covers the following: (1) data preparation, cleansing and selection; (2) data processing and analyses, and (3) knowledge representation and visualization (Fig. 2). The adaptability of the software architecture for knowledge discovery and decision making is accomplished by scalable experimental framework supporting: (1) various models and methods applied; (2) scalability; and (3) polymorphic computational architecture.

The designed architecture contains of layers for searching and integration of heterogeneous data from different data sources and in various formats; data preparation, cleansing, filtering and selection; data processing and analyses, and knowledge representation and results visualization.

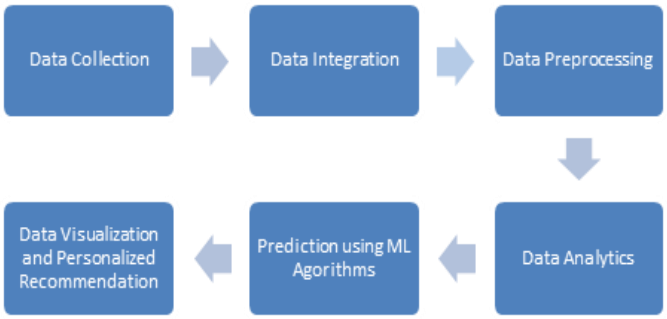

Fig. 2 Knowledge discovery from big biomedical data workflow process

Data Aggregation Layer applies methods for retrieving, structuring and storing large volumes of data collected from different sources – clinical data, personal data, genetic data, and medical images. Searching and integration services allow rapid management of large diverse data sets represented in different formats - relational, NoSQL, flat files. The integration system consists of services for integration of data in different format and from various sources, transforming the common request into a specific language request for each local database, depending on its type.

Processing and Analytics Layer performs the processing and analysis to make predictive modeling based on the data collected in the data storage. Scalable methods and algorithms for data analysis and the presentation of open knowledge in the field of artificial intelligence such as machine learning, models based on social behavior, metaheuristic, neural networks, and topological analysis are used as basic methods for data analytics as well as models for 3D stereoscopic visualization for better understanding of knowledge discovered.

B. Approbation

The proposed software framework has been verified and validated for the following case studies: distributed biological data searching and integration [18], management of large amount of heterogeneous data sets from various data sources for a breast cancer diagnostic system [19], gene sequences alignment [20], and three-dimensional visualization of the proteins structure or DNA sequences [21].

On the basis of research and comparative analysis, for each target scientific area of the study spectrum, and with the

expertise of research scientists, a specific approach is determined for the selection of basic analysis data and relevant attributes. Data identification and data abstraction play a crucial role in the data integration process. SOA based multiagent approach for distributed biological data searching and integration is aimed to automate the data integration and allows rapid management of large volumes of diverse data sets represented in different formats from different sources [18]. The system allows the user to set search criteria and access multiple databases simultaneously. The services allow access to the system via Internet by multiple clients (mobile phones, web browsers, desktop applications) and simultaneously serve a wide range of users.

An approach to management of large amount of heterogeneous data sets from various data sources for a breast cancer diagnostic system is presented [19]. Big genomic data architecture consists of data sources, storage, integration and preprocessing, real data stream, stream processing, analytical data store, analysis and reporting. Activities at data management for breast cancer diagnostic system are explained. Conceptual database architecture for storing data sets of several types in order to support breast cancer prediction is designed. The breast cancer database comprises of information related to breast cancer genes and functions - id, name, type, organism, function, and proteins coded, description, link for retrieving sequence. The patient's database consists of individual patient data - genetic data, clinical history, individual life style parameters, clinical tests results, environmental factors. The data sets in the suggested big data management system are retrieved from the biomedical research databases. The data management system is platform independent, easy to use and provides access to other databases such PubMed, NCBI. The purpose is to be used for data storage in a system for big data analytics and knowledge discovery, especially for the case study of breast cancer diagnostic. The advantages in data management, analysis, and knowledge discovery empower the scientists to achieve new scientific breakthroughs. As a result the research work is directed towards rapid management and processing of clinical data for solving problems in the field of precision medicine.

An innovative effective and unified method for DNA sequence alignment based on the trilateration, called CAT method, is designed and implemented on .NET platform using the C# programming language [20]. This method suggests solutions to three major problems in sequence alignment: creating a constant favorite sequence, reducing the number of comparisons with the favorite sequence, and unifying / standardizing the favorite sequence by defining benchmark sequences, which allow making comparisons at the outset – during input of the sequences in the database and it can be stored as meta data to each sequence.

A software application for biological data visualization has been developed for the purpose of system testing and validation [21]. The proposed application provides an opportunity for three-dimensional visualization of the proteins structure or DNA sequences. The three-dimensional modeling of the corresponding macromolecules enables one to gain a clear view of the objects complexity at the atomic level. Complex molecules can be displayed by using modern technologies for 3D modeling.

IV. Breast Cancer Data Classification Based on Machine Learning

Algorithm for Breast Cancer Prediction Based on Machine Learning

The breast cancer prediction algorithm based on machine learning is presented in Fig. 3.

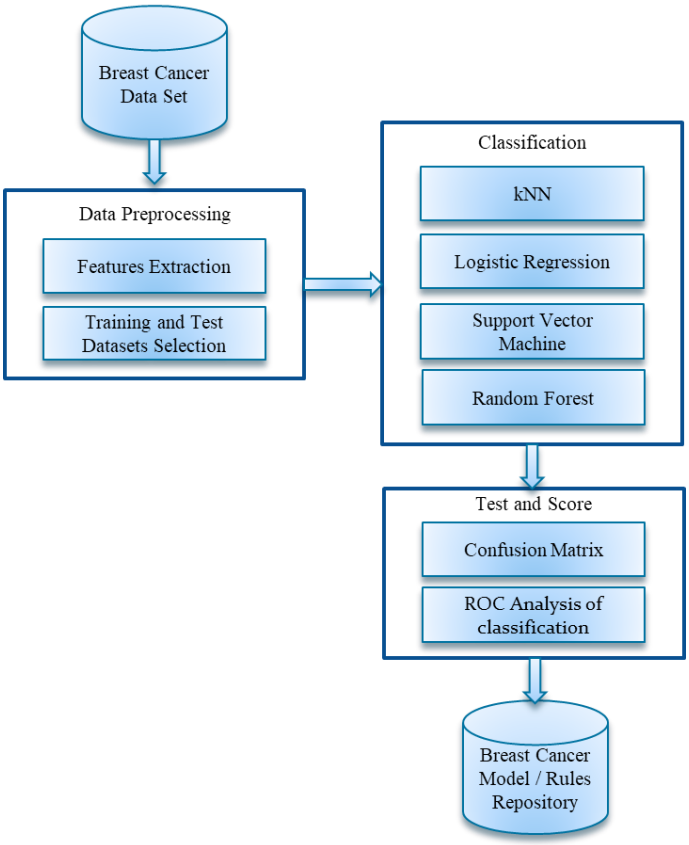

Fig. 3 Algorithm for breast cancer prediction based on machine learning

The research techniques follow the processing pipeline for discovering useful knowledge from a collection of data and cover the following: data preparation, cleansing and selection; knowledge discovery and decision making, and comprising results and interpreting accurate solutions from the observed results. The responsibilities of the preprocessing phase is preprocessing of the training and testing data sets. Preprocessing of data in knowledge discovery covers: data clearing in terms of accuracy; selection of functions in terms of relevance and features extraction. Features selection is very important as it contains information that can be used to train the system for identifying specific patterns. The aim is to establish model repository using training and testing data sets,

and applying various machine learning algorithms for classification. The second step is an analysis of all these features for detecting and classifying a possible pattern. Finally, the step involves a ML algorithm to determine the most appropriate model to represent the behavior or the pattern of the data. ML phase performs offline on the training and validation data sets, and is based on various methods for classification. Analytical model is created after execution of the feature extraction and dataset reduction process. As a result, various classification models are created and are used to build analytics workflow. The purpose of the ML phase is to build up models and rules repository used in the prediction phase.

Data Set Selection

Diagnostic Wisconsin Breast Cancer data set have been obtained from the UCI machine learning repository, which is available through open access and used for case study of differentiates benign (non-cancerous) and malignant (cancerous) samples [22]. The data set consists of records collected from biopsies of real patients with malignant or benign tumor type in various hospitals in Wisconsin and is grouped chronologically in the order in which the original clinical cases were reported. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

The data set contains of 569 instances or samples characterized by attributes as following:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The average, standard error and worst (average of the three highest values) of these features were computed for each image, resulting in 30 features.

The goal is to classify whether the breast cancer is benign or malignant. To achieve this machine learning classification methods are used to fit a function that can predict the discrete class of new input.

The example view of original data set is shown in Fig. 4, where each instance is represented by a row associated with the attributes value.

"id","diagnosis","radius_mean","texture_mean","perimeter_mean","area_mean","smoothness_mean","compactness_mean","concavity_mean","concave points_mean","symmetry_mean","fractal_dimension_mean","radius_se","texture_se","perimeter_se","area_se","smoothness_se","compactness_se","concavity_se","concavepoints_se","symmetry_se","fractal_dimension_se","radius_worst","texture_worst","perimeter_worst","area_worst","smoothness_worst","compactness_worst","concavity_worst","concave points_worst","symmetry_worst","fractal_dimension_worst",

842302,M,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,153.4,0.006399,0.04904,0.05373,0.01587,0.03003,0.006193,25.38,17.33,184.6,2019,0.1622,0.6656,0.7119,0.2654,0.4601,0.1189

842517,M,20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05667,0.5435,0.7339,3.398,74.08,0.005225,0.01308,0.0186,0.0134,0.01389,0.003532,24.99,23.41,158.8,1956,0.1238,0.1866,0.2416,0.186,0.275,0.08902

84300903,M,19.69,21.25,130,1203,0.1096,0.1599,0.1974,0.1279,0.2069,0.05999,0.7456,0.7869,4.585,94.03,0.00615,0.04006,0.03832,0.02058,0.0225,0.004571,23.57,25.53,152.5,1709,0.1444,0.4245,0.4504,0.243,0.3613,0.08758

Fig. 4 Example of original data set

Data Processing

The training phase is aimed to establish model repository using training data set, and applying classification machine learning algorithms. Analytical models are created after execution of the feature extraction process.

The attribute Diagnosis is selected as target for the classification. The attributes ID is selected as meta data in order to exclude it from the classification process since it has no relevance to features attribute value. As training data set are selected 376 samples, i.e. 66 %. Remaining 193 samples are used as test data set.

The proposed framework will be used for different machine learning algorithms investigation. In order to validate this approach four machine learning algorithms for classification are selected and used: Random Forest, kNN, Logistic Regression, and SVM. The experiment is implemented as workflow using Orange Data Mining tool [23]. The workflow for breast cancer data classification is presented in Fig. 5.

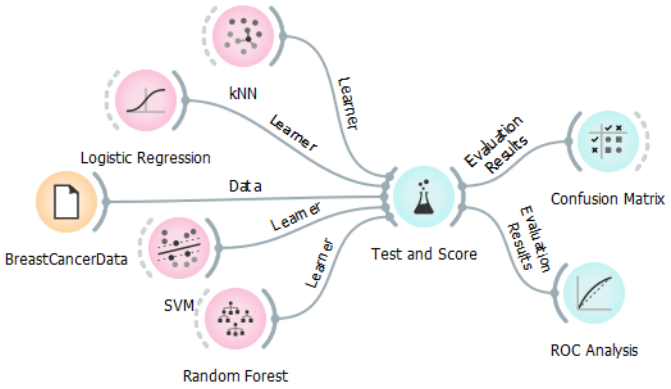

Fig. 5 Workflow for breast cancer data classification

Random Forest classification method consists of a large number of individual decision trees and is "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [24]. The second machine learning algorithm is k-nearest neighbors (kNN) and it was selected which "finds a group of k objects in the training

set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood" [25]. Logistic Regression is used to assign observations to a discrete set of classes and it based on the concept of probability [25]. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1.

The last machine learning algorithm is Support Vector Machines (SVM) which requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, SVM is efficient methods for training [25]. SVM find the best classification function to distinguish between members of the two classes in the training data. Once this function is determined, new data instance can be classified and belongs to the positive class.

V. Experimental Results and Analysis

Precision is one of evaluation metrics of the model performance and is calculated as a ratio of true positive classified items divided by sum of true positive and false positive items in the test set. The precision range is from 0 (least precision) to 1 (most precision). The measured results for precision obtained from the selected classification algorithms are presented in Fig. 6. Best result in terms of precision is achieved through SVM classification algorithm: 0.977.

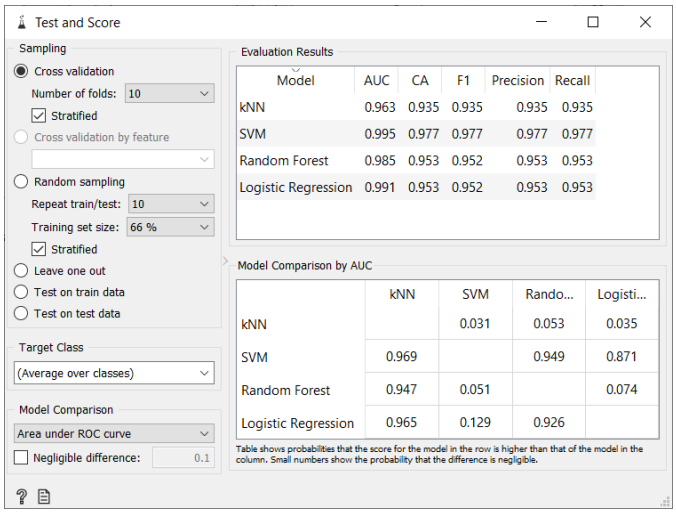

Fig. 6 Test and score of selected classification algorithms

ROC curves are used to observe the classifiers and comparison between classification models. ROC curves for the tested models and results of testing classification algorithms are presented in Fig. 7. ROC curve demonstrates several things: It shows a trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left border and then the upper border of the ROC space, the more accurate the test is. The curve plots a false positive rate on an x-axis against a true positive rate on a y-axis. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the classifier. Given the costs of false positives and false negatives, it can be also determined the optimal classifier, in this case that is Random Forest.

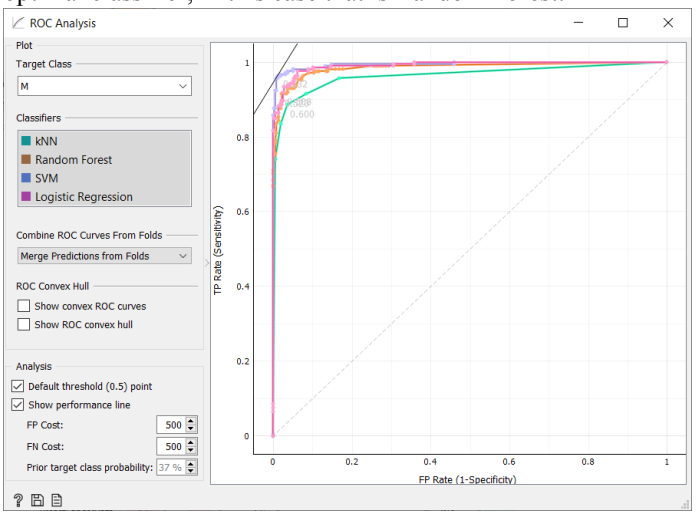

Fig. 7 ROC analysis of the classification models

The number of instances between the actual and the predicted class are presented in confusion matrix (Fig. 8). The matrix is useful for monitoring which specific cases are misclassified.

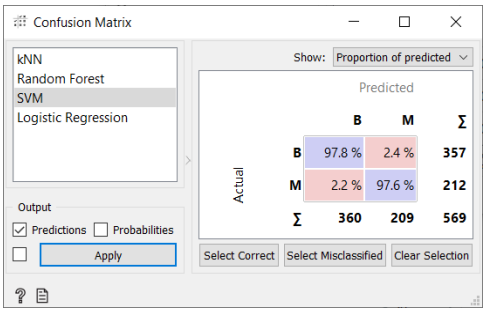

Fig. 8 Confusion matrix for case of SVM algorithm

VI. Conclusion

A platform for multidimensional large-scale biomedical data management and analytics is presented in this paper. The suggested platform aims intelligent data management, analysis and visualization. The advantages in data management, analysis, knowledge discovery and visualization empower the scientists to achieve new scientific breakthroughs. As a result the research work is directed towards developing computer aided diagnostic system for solving problems in the field of precision medicine.

A breast cancer prediction algorithm based on machine learning is presented. The research techniques follow the processing pipeline for discovering useful knowledge from a collection of data and cover the following: data preprocessing; knowledge discovery and decision making; comprising results and interpreting accurate solutions from the observed results. Four machine learning algorithms for breast cancer classification are selected and evaluated experimentally: Random Forest, kNN, Logistic Regression, and SVM.

ACKNOWLEDGMENT

This paper presents the outcomes of a research project "Innovative Platform for Intelligent Management and Analysis of Big Data Streams Supporting Biomedical Scientific Research", KP-06-N37/24, funded by the National Science Fund, Ministry of Education and Science, Bulgaria.

References

[1] A. Labrinidis and H. Jagadish, "Challenges and opportunities with big data," in *Proceedings of the VLDB Endowment,* 5(12):2032-2033, DOI: 10.14778/2367502.2367572.

[2] T. Hulsen, et all, "From big data to precision medicine," *Front. Med.,* March 2019, DOI: https://doi.org/10.3389/fmed.2019.00034.

[3] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Inf Sci Syst* 2, 3, 2014, doi: 10.1186/2047-2501-2-3.

[4] P. Chen, C. Zhang "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Journal of Information Sciences*, Vol. 275, 2014, pp. 314-347, https://doi.org/10.1016/j.ins.2014.01.015.

[5] X. Jin, B. W. Waha, X. Cheng, Y. Wang "Significance and challenges of big data research," *Big Data Research,* Vol. 2, Issue 2, June 2015, pp. 59-64, https://doi.org/10.1016/j.bdr.2015.01.006.

[6] Y.-F. Lu, D. B. Goldstein, M. Angrist and G. Cavalleri, "Personalized medicine and human genetic diversity," *Cold Spring Harb Perspect Med.,* 2014 Sep; 4(9): a008581. doi: 10.1101/cshperspect.a008581.

[7] G. S. Ginsburg and K. A. Phillips, "Precision medicine: from science to value," *Health Aff (Millwood),* 2018 May; 37(5): 694–701. doi: 10.1377/hlthaff.2017.1624.

[8] A. Abed, J. Yuan and L. Li, "Based SVM distinct stages framework data mining technique approach for text extraction," *WSEAS Transactions on Information Science and Applications,* pp.100-110, Volume 16, 2019.

[9] G. S. Ow and V. A. Kuznetsov, "Big genomics and clinical data analytics strategies for precision cancer prognosis," *Scientific Reports* 6, Article 36493, 2016, https://doi.org/10.1038/srep36493.

[10] M. Panahiazar, V. Taslimitehrani, A. Jadhav and J. Pathak, "Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases," in *Proc IEEE Int Conf Big Data,* 790–795, 2014, doi: 10.1109/BigData.2014.7004307.

[11] M. Viceconti, P. Hunter and R. Hose, "Big data, big knowledge: big data for personalized healthcare," *IEEE J Biomed Health Inform* 19, 1209–1215, 2015, doi: 10.1109/JBHI.2015.2406883.

[12] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput Struct Biotechnol J.* 2015; 13: 8–17, doi: 10.1016/j.csbj.2014.11.005.

[13] European Institute of Bioinformatics, https://www.ebi.ac.uk/about/our-impact.

[14] R. Luo, et all, "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." *Gigascience*. 2012 Dec 27;1(1):18. doi: 10.1186/2047-217X-1-18.

[15] Y. Li and L. Chen, "Big Biological Data: Challenges and Opportunities." *Genomics Proteomics Bioinformatics*. 2014 Oct; 12(5): 187–189. doi: 10.1016/j.gpb.2014.10.001

[16] NVIDIA Clara: Healthcare application framework for AI-powered imaging and genomics, https://developer.nvidia.com/clara.

[17] Caffe: Deep learning framework, https://caffe.berkeleyvision.org/.

[18] V. Gancheva, "SOA based multi-agent approach for biological data searching and integration," *International Journal of Biology and Biomedical Engineering*, ISSN: 1998-4510, Vol. 13, 2019, pp. 32-37.

[19] V. Gancheva, "A big data management approach for computer aided breast cancer diagnostic system supporting precision medicine." *AIP Conference Proceedings*, 2172, 090012 (2019); https://doi.org/10.1063/1.5133589.

[20] V. Gancheva and H. Stoev, "DNA sequence alignment method based on trilateration," In: Rojas I., Valenzuela O., Rojas F., Ortuño F. (eds) Bioinformatics and Biomedical Engineering, IWBBIO 2019, *Lecture Notes in Computer Science*, vol. 11466, Springer, Cham, pp. 271-283, https://doi.org/10.1007/978-3-030-17935-9_25.

[21] V. Gancheva "Knowledge discovery based on data analytics and visualization supporting precision medicine," in *Proc of International Conference on Mathematics and Computers in Science and Engineering* 2020, DOI 10.1109/MACISE49704.2020.00024.

[22] W. H. Wolberg, Breast Cancer Wisconsin (Original) Data Set, University of Wisconsin Hospitals, Madison, Wisconsin, USA, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

[23] Orange Data Mining, [Online]. Available: https://orange.biolab.si/

[24] L. Breiman, "*Random Forests*." Machine Learning. 45 (1): 5–32. doi:10.1023/A:1010933404324.

[25] Wu, Xindong, et al. "*Top 10 algorithms in data mining*." Knowledge and information systems 14.1 (2008): 1-37.

Veska Gancheva, Asoc. Prof. PhD is scientist and university professor of the Technical University of Sofia. She is expert with experience in information technology, bioinformatics, in silico biological experiments, parallel methods, algorithms and models, software technologies, supercomputing applications, cloud computing, big data management and analysis. Asoc. Prof. Veska Gancheva is author of 70 scientific publications at international scientific conferences and journals and has participated in over 30 research projects at national and European level: Seventh Framework Programme, PRACE, ERASMUS+, and the Operational Programme for Human Resources Development.