МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2024
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2024
Proceedings of the Fifty-Third Spring Conference
of the Union of Bulgarian Mathematicians
Borovets, April 1–5, 2024

BIG BIOMEDICAL DATA ANALYTICS IN SUPPORT OF PRECISION MEDICINE*

Desislava Ivanova
Technical University of Sofia, Faculty of Applied Mathematics and Informatics,
Sofia, Bulgaria
e-mail: d_ivanova@tu-sofia.bg

The goal of this paper is to present the major outcomes of the author scientific research in the area of Big Biomedical Data analytics in support of precision medicine, specifically genomics and oncology. The main parts of the paper reveal the state-of-the art in big data technologies and their significance for the advance of molecular biology and medicine in the context of the computational paradigm for scientific research. Furthermore, it presents approaches for gene finding and early detection of breast cancer and thyroid cancer based on big biomedical data analytics.
Keywords: big data, biomedical analysis, knowledge data discovery, precision medicine.

АНАЛИЗ НА ГОЛЕМИ БИОМЕДИЦИНСКИ ДАННИ В ПОДКРЕПА НА ПРЕЦИЗНАТА МЕДИЦИНА

Десислава Иванова
Технически университет – София, Факултет Приложна математика и информатика,
София, България
e-mail: d_ivanova@tu-sofia.bg

Целта на тази статия е да представи основните резултати от научните изследвания на автора в областта на анализа на големи биомедицински данни в подкрепа на прецизната медицина, по-специално геномиката и онкологията. Основните части на статията разкриват най-съвременните технологии за големи данни и тяхното значение за напредъка на молекулярната биология и медицина в контекста на изчислителната парадигма за научни изследвания. Освен това, статията представя подходи за откриване на гени и ранно откриване на рак на гърдата и рак на щитовидната жлеза въз основа на анализ на големи биомедицински данни.
Ключови думи: големи данни, биомедицински анализ, откриване на данни от знания, прецизна медицина.

*https://doi.org/10.55630/mem.2024.53.018-024
2020 Mathematics Subject Classification: 68-XX, 92-XX.

1. Introduction. In recent times, Big Data has garnered widespread recognition as a transformative force in scientific research and a highly promising trend in the realm of Information Technology (IT). This paradigm shift, labelled "Data-Intensive-Scientific-Discovery" (DISD), is reshaping the landscape of experiments and knowledge discovery. Advanced technologies like Big Data Analytics, Internet of Things (IoT), and Cloud Computing present researchers with potent tools for Knowledge Data Discovery (KDD) and intelligent decision-making. These KDD solutions play a pivotal role in detecting intricate DNA anomalies, identifying rare genetic diseases, and addressing cancer. This forward-thinking approach proves particularly advantageous in the analysis of big biomedical data for precision medicine, offering substantial societal and health-related benefits. Precision medicine has been one of the hottest topics nowadays and involves disease treatment that considers individual genetic profile, environmental specifics, and lifestyle of the individual.

2. Big data in scientific research. In recent years, Big Data has been recognized by eminent scientists, researchers, and analysts as a transformative force in scientific studies and a challenging trend in Information Technology (IT) [4, 7]. This acknowledgment has accelerated the development of methods and technologies for processing large volumes of data, leading to profound shifts in scientific research paradigms. The emergence of the novel scientific paradigm, "Data-Intensive Scientific Discovery (DISD)", has revolutionized scientific research and innovations, encompassing phases such as data accumulation, filtering, integration and presentation, analysis, and data-intensive decision making. This paradigm introduces challenges in processing technologies, including data accumulation and storage, searching, sharing, analysis, visualization, high-performance processing resources, parallel and distributed processing, parallel input/output, and in-memory processing. Scalability and streaming pose primary challenges in Big Data analysis, revolutionizing fundamental scientific studies in molecular and computational biology.

In 2001, the formal definition of Big Data introduced the 3V model by Doug Laney, encapsulating Volume, Velocity, and Variety. Prominent entities like Gartner, IBM, and Microsoft continue to use this "3Vs" model to delineate Big Data characteristics. A 2011 IDC report asserts the emergence of new technologies to extract value from diverse and big data, leading to the 4V model—Volume, Variety, Velocity, and Value. The modernized 5V model introduces Veracity, emphasizing data accuracy and reliability within Big Data.

The Strategic Research and Innovation Agenda (SRIA) of the European Big Data Value Association (BDA) highlights Big Data's revolutionary impact on fundamental science, with biology and medicine leading the way. The next generation of data analysis methods must adeptly handle diverse big data sources, characterized by varying attributes, trust levels, and update frequencies. Effective knowledge acquisition through data analysis is crucial, as the European Commission designates Big Data as a primary asset for the development of fundamental scientific research across all sectors.

Molecular and computational biology, heavily reliant on big data, is a key area of focus, with intensive research in bioinformatics driving innovative methods for processing and analyzing biological and biomedical data.

3. Big data in support of precision medicine. It is widely recognized that a medical treatment effective for some patients may not be suitable for others, and indi-

viduals can experience varied reactions to the same drug. The aspiration to tailor medical approaches to individual characteristics has been a longstanding goal in medicine, with Precision Medicine referring to the customization of medical treatment based on individual patient characteristics. This involves categorizing individuals into subpopulations with variations in disease sensitivity, biology, prognosis, or response to specific treatments, rather than creating unique drugs or medical devices for each patient. In cancer treatment, understanding the genetic profile is crucial, especially for diseases like breast cancer considered genetic disorders due to mutations in specific genes. Precision medicine groups patients based on their specific characteristics, facilitating the assignment of optimal treatment based on the characteristics of the target group [5].

The journey of precision medicine begins with genomics, relying on omics platforms for comprehensive analysis and multi-scale data interpretation. Big Genomic Data platforms like Google Genomics, IBM Reference Architecture for Genomics, and SAP®Connected Health platform serve as collaborative hubs, fostering the creation of patient-centric solutions to enhance healthcare, reduce costs, and offer connected healthcare services. Since the year 2000, the medical and biological sciences have entered the post-genomics era, marked by the advent and significant development of genomics. The rise of genomics is driven by advanced technology for complete genomic sequencing, specifically determining the nucleotide sequence of entire genomes [1, 5].

The big genomic data ecosystem's conceptual model includes the latest DNA sequencing technologies, the generation of "omics" data, in silico technologies producing insilico experimental data, genome databases, cancer genome databases, and associated technologies like the Internet of Medical Things (IomT) and cloud technologies. A new challenge has emerged as the capacity of genomic databases is expanding faster than the capability of analytical tools.

In addition to genomics, medical images represent a rapidly expanding source of Big biomedical data in healthcare. The volume of storage required for medical images has tripled since 2005, prompting healthcare institutions to turn to cloud solutions for ample storage capacity. Medical imaging cloud technologies encompass web-based platforms for imaging analytics, offering high-speed cloud computing infrastructures, advanced visualization, deep learning, and cloud-based storage and sharing of medical images. The collection and aggregation of patient-generated health data from Internet of Medical Things (IomT) devices play a pivotal role in population health management, emphasizing the importance of proficiency in Big Data Analytics for society to deliver high-quality patient care.

4. Big genomic data analytics for gene finding. In the current landscape, there is a critical need for efficient algorithms and computational power to analyse the escalating volumes of data across diverse industries. Genomics emerges as a prominent sector within the realm of big data, requiring not only the extraction of meaningful information but also the acquisition of knowledge, uncovering insights, identifying patterns, and making sense of the data. Past methods, including statistical and graph theoretical approaches, data mining, and machine learning, have seen partial success in terms of performance, particularly in dealing with the complex nature of living organisms, where the coordinated actions of different gene groups play a crucial role. Identifying the "active and deactivated" genes, especially in the genomics era, poses a significant challenge, with promoters being key elements involved in differential transcription regulation [6].

Promoters, situated near the starting point of protein biosynthesis, play a pivotal role in coordinating gene expression through distinct transcriptional control mechanisms. The presented results stem from experimental exploration focused on detecting enhancer-promoter interactions in Genomic Big Data, utilizing machine learning algorithms such as Decision Tree (DT) and Support Vector Machine (SVM) [8]. The pipeline involves pre-processing, classifier application to identify enhancer-promoter interactions, and presenting the experimental results. Feature extraction from DNA sequences of enhancer and promoter elements is conducted using two approaches: searching for known transcription factor binding site motifs and employing a word embedding model for direct sequence embedding into a new feature space. The subsequent stage focuses on data classification based on extracted features, employing Decision Tree and Support Vector Machine classifiers.

Operating within the Apache Spark environment [12], the experimental framework facilitates streaming and real-time analysis of big data. Leveraging the Apache Spark machine learning library (MLlib) in Python, the investigation utilizes enhancer-promoter interaction data from the TargetFinder project [14], encompassing six cell lines. The developed software, coded in Python, implements the proposed approach using Decision Tree and Support Vector Machine classifiers. The detection of enhancer-promoter interactions, a challenging genomics phenomenon, is experimentally explored through machine learning, demonstrating competitive accuracy levels ranging from 91% to 95%.

5. Big data in support of cancer detection. Today, medical treatment plans are often generic and not specifically tailored to individual patients, resembling what doctors would prescribe to anyone with the same condition. This approach is rooted in "standards of care", which are the best courses of prevention or treatment for the general population. In the context of breast cancer, these standards typically involve routine self-exams, mammograms after a certain age, and conventional chemotherapy if a tumor is detected. In case the initial treatment proves ineffective, doctors and patients proceed to the next option in a trial-and-error fashion, with significant stakes involved [1, 5].

Each person possesses a unique variation in the human genome, and while most variations have no impact on health, an individual's well-being is influenced by genetic variations along with behaviours and environmental factors. Modern advances in personalized medicine leverage technology to confirm a patient's fundamental biology, including DNA, RNA, or proteins, ultimately leading to the confirmation of diseases [4]. The concept of personalized medicine can be applied to innovative and transformative approaches in healthcare, enabling predictions of a person's risk for a particular disease based on one or several genes. All these activities aligned with the concept of personalized medicine generate a substantial amount of data (big medical datasets) that requires processing and analysis.

5.1. Big data analytics for early detection of breast cancer. Breast cancer initiates when breast cells undergo uncontrolled growth, resulting in a tumor that can be detected through an X-ray or felt as a lump. The malignancy of the tumor is determined by its ability to invade surrounding tissues or metastasize to distant areas of the body. While predominantly affecting women, breast cancer can also impact men, originating from different parts of the breast, often starting in the ducts or glands responsible for milk production [2, 4].

Personalized medicine plays a significant role in advancing preventive care for breast cancer. For example, women with a family history of breast or ovarian cancer are often genotyped for mutations in the BRCA1 and BRCA2 genes. These genes produce tumor suppressor proteins crucial for DNA repair and stability. Mutations in these genes, when inherited, elevate the risk of several cancers, including breast and ovarian cancer. The proposed conceptual model for early breast cancer detection aims to provide expert oncologists with a second opinion, enhancing accuracy, reducing biopsy rates, and saving time. The model involves stages such as data cleaning, feature extraction, classification using the Naive Bayes classifier, and performance analysis, achieving an accuracy of 0.9786 on the Wisconsin breast cancer database [9].

In global breast cancer research, various techniques are employed to process medical data for identification, presenting challenges due to dataset readability, diverse structures, and consideration of different attributes. The creation of benchmark breast cancer datasets is a pressing future objective, facilitating the evaluation and comparison of algorithmic results. Collaborative efforts between IT specialists and oncology experts are crucial to categorize breast cancer anomalies accurately and establish relevant datasets for comprehensive analysis.

5.2. Big data analytics for Early detection of thyroid cancer. The thyroid gland, situated in the lower neck and resembling a butterfly, plays a pivotal role in metabolism regulation by releasing hormones like Triiodothyronine (T3) and Thyroxine (T4). Discrepancies in hormone production, along with the presence of nodules or tumors (benign or malignant), can lead to thyroid disorders. Detection methods involve physical examinations, blood tests for hormone levels, and ultrasound for confirming the presence of thyroid nodules.

Medical imaging has undergone significant transformations, encompassing techniques like X-ray radiography, computed tomography (CT), and ultrasound. Ultrasound imaging, utilizing sound waves, stands out for its real-time images, portability, cost-effectiveness, and absence of harmful radiation, making it a primary method for thyroid anomaly detection [10].

Recent years have witnessed extensive research in thyroid cancer detection, facilitated by tools and databases developed through computer technology. Innovations like EU-TIRADS classification, Cancer Genome Atlas (CGA), and Thyroid Cancer Risk Assessment Tool (TCRA) have streamlined research, enhancing thyroid cancer early detection and personalized therapy [3, 13].

The early thyroid cancer detection approach utilizes ultrasound medical image analysis is proposed involving pre-processing, anomaly detection, and post-processing stages. The pre-processing phase eliminates noise and performs attribute selection. Anomaly detection integrates image segmentation using Convolutional Neural Networks (CNNs) and Support Vector Machines (SVM) for classification. The post-processing stage presents detected anomalies for interpretation by thyroid cancer experts.

Implemented in Apache Spark with MLlib [12], the conceptual model utilizes CNNs and SVM for classification. CNNs, with five layers and 500 nodes per convolution layer, enhance accuracy. SVM, known for handling large datasets, uses a kernel function for optimal hyperplane selection in a higher-dimensional space [11, 12]. Medical imaging analysis reveals SVM outperforms CNN, achieving 87.8% accuracy compared to CNN's 72.9%. SVM exhibits greater stability, although CNNs consume more time for classifica-

tion with exponentially increasing training time based on input data dimension.

6. Conclusion. This paper presents some results and experience of the scientific research in the interdisciplinary area of high IT in support of precision medicine. It has taken years and great efforts of collaboration with our partner experts in molecular biology and medicine to conduct our investigation.

The major goal has been to build up intelligent digital solutions based on Big data technologies to help medical experts and researches in molecular biology process the huge amount of information associated with their analysis and decisions and basically, to support them in extracting knowledge (value) effectively out of the relevant Big biomedical data. The results of the scientific projects have been successfully applied in a number of our BSc and MSc courses at the Technical University of Sofia such as Bioinformatics, Medical Informatics, IoT Ecosystem Security, Big Data Analytics, Big Data Technologies in Support of Precision Medicine, Programming for IoT Ecosystem, etc. The future work implies porting our integrated digital solution in support of precision medicine on high performance computer system, the focus being on streaming and GPU accelerated infrastructures.

REFERENCES

[1] A. Akalin, M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick, C. E. Mason. Methylkit: a comprehensive R package for the analysis of genome-wide DNa methylation profile. *Genome Biol.*, **13**, (2012), R87 `https://doi.org/10.1186/gb-2012-13-10-r87`.

[2] Breast Cancer Surveillance Consortium, Risk Estimation Datasets & Risk Factors Dataset:`https://www.bcsc-research.org/data`.

[3] Cancer Genome atlas, `cancergenome.nih.gov/`.

[4] Challenges and opportunities with Big Data, a community white paper available at: `http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf`.

[5] P. Galetsi, K. Katsaliaki, S. Kumar. Big data analytics in health sector: Theoretical framework, techniques and prospects. *Int. J. Inf. Manage.*, **50** (2020), 206–216.

[6] eNCoDe (encyclopedia of DNa elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI), `https://www.encodeproject.org/`.

[7] R. H. Hariri, E. M. Fredericks, K. M. Bowers. Uncertainty in big data analytics: survey, opportunities, and challenges. *J. Big Data*, **6** (2019) 44.

[8] D. Ivanova, P. Borovska, V. Gancheva. Experimental Investigation of Enhancer-Promoter Interactions out of Genomic Big Data based on Machine Learning. *International Journal of Computers*, **3** (2018), 58–62, `http://www.iaras.org/iaras/journals/ijc`.

[9] D. Ivanova. Big Data analytics for early Detection of Breast Cancer Based on Machine Learning. Proceedings of the 43rd International Conference applications of Mathematics in engineering and economics, aIP Conf. Proc. 1910, `https://doi.org/10.1063/1.5014010`.

[10] D. Ivanova. Artificial intelligence in internet of medical imaging things: The power of thyroid cancer detection, Scopus. 2018 International Conference on Information Technologies, InfoTech 2018 Proceedings 8510725;, IEEE, DOI: 10.1109/InfoTech.2018.8510725.

[11] D. Ivanova. Internet of Medical Imaging Things and the Application of Information Technologies for Early Detection of Thyroid Cancer. *Scientific Journal "In-Silico Intellect"*, **1**, no. 1 (2017), 20–25.

[12] MLlib apache Spark: `https://spark.apache.org/mllib`.

[13] W. H. Wolberg. Breast Cancer Database, University of Wisconsin Hospitals, Madison, Wisconsin, USA.
[14] TargetFinder project: `https://github.com/carringtonlab/TargetFinder`.