Journal of Big Data

RESEARCH Open Access

Big Data in multiscale modelling: from medical image processing to personalized models

Tijana Geroski[1,2], Djordje Jakovljević[3] and Nenad Filipović[1,2*]

*Correspondence:
fica@kg.ac.rs

[1] Faculty of Engineering, University of Kragujevac, Kragujevac, Serbia
[2] Bioengineering Research and Development Centre (BioIRC), Kragujevac, Serbia
[3] Coventry University, Coventry, UK

Abstract

The healthcare industry is different from other industries–patient data are sensitive, their storage needs to be handled with care and in compliance with regulative, while prediction accuracy needs to be high. This fast expansion in medical image modalities and data collection leads to generation of so called "Big Data" which is time-consuming to be analyzed by medical experts. This paper provides an insight into the Big Data from the aspect of its role in multiscale modelling. Special attention is paid to the workflow, starting from medical image processing all the way to creation of personalized models and their analysis. A review of literature regarding Big Data in healthcare is provided and two proposed solutions are described–carotid artery ultrasound image processing and 3D reconstruction, and drug testing on personalized heart models. Related to the carotid artery ultrasound image processing, the starting point is ultrasound images, which are segmented using convolutional neural network U-net, while segmented masks were further used in 3D reconstruction of geometry. Related to the drug testing on personalized heart model, similar approach was proposed, images were used in creation of personalized 3D geometrical model that is used in computational modelling to determine pressure in the left ventricle before and after drug testing. All the aforementioned methodologies are complex, include Big Data analysis and should be performed using servers or high-performance computing. Future development of Big Data applications in healthcare domains offers a lot of potential due to new data standards, rapid development of research and technology, as well as strong government incentives.

Keywords: Big Data, Multiscale modelling, Medical image processing, 3D reconstruction

Introduction

The term "Big Data" has become a buzzword in recent years, as the frequency of its use has doubled every year over the previous decade (Andreu-Perez et al. 2015). Big Data is described by three primary features known as the "3V": volume (the quantity of data created), variety (data from many categories), and velocity (the rate at which data is generated) (Luo et al. 2016); (Viceconti et al. 2015); (Oussous et al. 2018); (Belle, et al., 2015); (Yang et al. 2017). Recently, additional two "Vs" have been introduced: variability (data

inconsistency) and veracity (quality of recorded data) (Viceconti et al. 2015); (Yang et al. 2017). As a result, the "5V" now identifies huge data concerns (Kouanou et al. 2018).

Big Data applications are used in many disciplines of research, including agriculture (Wolfert et al. 2017); (Zhang et al. 2015), internet with social networks (Pääkkönen and Pakkala 2015), as well as medicine (Andreu-Perez et al. 2015); (Luo et al. 2016); (Viceconti et al. 2015); (Belle et al. 2015) and personalized medicine based on genomics data (Cirillo and Valencia 2019) etc. Data volume in the medical area is expanding, and traditional methodologies cannot handle such amounts adequately. Management, analysis, and storage of biological data are ongoing challenges in biomedical computing. As a result, Big Data technologies contain new frameworks for processing medical data playing an important role in data management, organizing, and analysis through the use of machine learning and deep learning approaches (Kouanou et al. 2018). It also enables fast data access via the NoSQL database (Kouanou et al. 2018). In the area of medical image analysis, due to significant improvement in image collecting equipment, the data is relatively huge (going to Big Data), which makes image analysis challenging (Razzak et al. 2018). It is said that due to digitalization of medical repositories in hospitals, as well as the use of medical images, digital medical archives size is growing at exponential rate (Ashraf et al. 2020a, b). According to a McKinsey Global Institute, if US healthcare uses Big Data creatively and efficiently, the sector could generate more than $300 billion in value per year. Two-thirds of the value would be realized through lowering US healthcare spending (Belle et al. 2015). This fast expansion in medical imagery and modalities necessitates considerable and time-consuming efforts by medical experts, who are subjective, prone to human error, as well as there are interpersonal differences. Using machine learning techniques to automate the diagnosis process is an alternative response to aforementioned challenges; however, typical machine learning methods are unable to cope with complex problems (Razzak et al. 2018). The successful combination of high-speed computers with machine learning promises the ability to cope with large amounts of medical image data for accurate and fast diagnosis (Razzak et al. 2018). In recent years, machine learning (ML) and artificial intelligence (AI) have advanced quickly, finding their tole in medical image processing, computer-aided diagnosis, image fusion, registration, image segmentation, as well as image-guided treatment. ML techniques extract information (called features) from images and effectively perform decision making (Razzak et al. 2018).

Problem definition

Big data in health refers to relevant datasets that are large, time-consuming and complicated for healthcare practitioners to manage and process using current technologies (Andreu-Perez et al. 2015); (Wang et al. 2017). Data is created at an unprecedented rate on a daily basis from several heterogeneous sources (e.g., laboratory and clinical data, patients' symptoms uploaded via remote sensors, hospital activities, and pharmaceutical data) (Oussous et al. 2018). As a result, new challenges have arisen such as storing, collecting, and interpreting vast volumes of data (Margolis et al. 2014). Techniques for biomedical imaging that are widely established in clinical settings computed tomography (CT), magnetic resonance imaging (MRI), X-ray, molecular imaging, ultrasound, photo-acoustic imaging, fluoroscopy, and positron emission

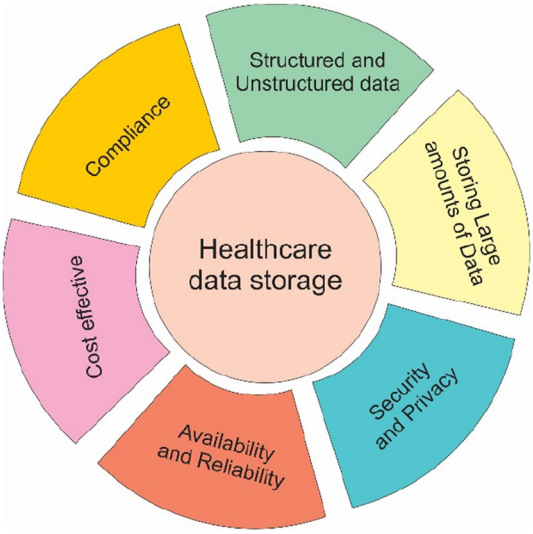

Fig. 1 Healthcare data storage

tomography (PET-CT) (Belle et al. 2015). These approaches provide high-definition medical images in vast amounts. However, doctors cannot diagnose all of the millions of images produced. With the increased availability of biomedical imaging data, additional demands are placed on Artificial Intelligence (AI) for machine learning (ML) systems to build complicated models. ML is here the fundamental mechanism used to extract structured information and knowledge from raw data and convert it into automatic predictions for a variety of applications (Xing et al. 2016). Imaging informatics in general is mostly used to improve the efficiency of image processing activities such storage, retrieval, and interoperation. In this section, we will focus on two important aspects: medical data storage and image processing in the area of Big Data.

Medical data storage in the era of big data

Medical image data can range from a few megabytes for a single study (for example, histology images) to hundreds of gigabytes per research [for example, thin-slice CT examinations with up to 2500+ scans per study (Seibert 2010)]. For example, the ImageCLEF medical image collection comprised around 66,000 images between 2005 and 2007, but in 2013, approximately 3,00,000 images were stored daily (Hersh et al. 2009). In order to adequately process large amounts of healthcare data, their storage needs to be in line with several aspects–compliance, security and privacy, cost effectiveness, availability and reliability (Fig. 1).

It is vital to explore how Big Data approaches (such as Hadoop and NoSQL databases) are utilized to store electronic health records (EHRs) (Luo et al. 2016). When working with clinical real-time stream data, effective data storage is critical (Dutta et al. 2011). Dutta et al. assessed the use of Hadoop and HBase as data warehouses for storing EEG data and addressed their high-performance qualities (George 2011). Sahoo et al. (2014) and Jayapandian et al. (2013) also presented a distributed system for storing and querying massive volumes of EEG data. Cloudwave, a system they designed, stores clinical data using Hadoop-based data processing modules, and includes web-based

interface for real-time data visualization and analysis by exploiting Hadoop's processing capabilities. Achieved results show tremendous save in processing time, Cloudwave was able to process five EEG studies in 1 min, in comparison to the stand-alone system that need 20 min for the same task. Jin et al. (2011) investigated the feasibility of employing Hadoop HDFS and HBase for distributed EHRs. In comparison to a standard relational database that is good at handling structured data, the innovative NoSQL is an advanced option for storing unstructured data. Mazurek (2014, May 27–30) presented a system that integrates relational and multidimensional technologies, as well as NoSQL repositories, to enable data mining techniques while also providing flexibility and speed in data processing. Nguyen et al. (2011) presented a prototype system for storing clinical signal data, in which clinical sensor time series data are stored within HBase in such a way that the row key serves as the time stamp of a single value, and the column stores patient physiological values that correspond with the row key time stamp. The information for the HBase data structure is stored in MongoDB62, a document-based database, to increase accessibility and readability.

The Picture Archiving and Communication System is one of the services used in medicine for image data storage and transfer (PACS) (Kouanou et al. 2018), which is achieved primarily through DICOM protocols in radiology departments (Luo et al. 2016). PACS are popular for transmitting images to local display workstations using existing protocols such as digital image communication in medicine (DICOM). However, data transmission with a PACS is very standardized (Doel et al. 2017), and this system depends entirely on structured data to retrieve medical images rather than utilizing the biomedical images' unstructured information (Istephan and Siadat 2016). To access PACS, several web-based medical apps have been developed, and growing use of Big Data technologies has improved their performance (Luo et al. 2016). Many studies have been conducted to manage and analyze structured and unstructured data images utilizing the Big Data and artificial intelligence concepts (Kouanou et al. 2018). Given the current tendency among health-care organizations to outsource the two critical components of PACS (DICOM object repository and database system) to the cloud, Silva et al. (2012) suggested a technique to integrate data in PACS. They suggested a Cloud input/output stream method in an abstract layer to accommodate many cloud providers no matter the differences in data access standards. Yao et al. (2014) built a huge Hadoop-based medical image retrieval system that retrieved the properties of medical images using a Brushlet transform and a local binary pattern algorithm, in addition to Big Data technologies based on the integration of cloud platforms with PACS. Image characteristics were then saved in HDFS, followed by MapReduce implementation. When compared to the findings without homomorphic filtering, the evaluation results showed a lower error rate. Similarly, Jai-Andaloussi et al. (2013) addressed the issues of content-based image retrieval systems using the MapReduce processing architecture and HDFS storage model. They performed testing on mammography datasets and achieved good results, demonstrating that the MapReduce approach may be utilized efficiently for content-based medical image retrieval.

Long-term storage of Big Data in medicine necessitates huge storage capacity. If any decision support system is to be conducted utilizing the data, it also necessitates quick and precise algorithms. Furthermore, if additional sources of data obtained for

each patient are also used during the diagnosis, prognosis, and treatment prediction, providing cohesive storage and designing effective systems capable of encapsulating the vast range of data becomes a difficulty (Belle et al. 2015). Compression methods can assist overcome data storage and network capacity limits when dealing with very large amounts of data. Many approaches for compressing medical images have been developed. However, a few approaches for large data compression have been devised. A technique for compressing both high-throughput sequencing datasets and data created by calculating log-odds of probability error for each nucleotide has been developed, with maximum compression ratios of 400 and 5, respectively (Ohno-Machado et al. 2012). Filtering and the Fourier transform were used as signal processing techniques in this model (Ohno-Machado et al. 2012). Wolff (2014) investigated the use of the simplicity and power (SP) theory of intelligence in massive data. SP theory seeks to simplify and combine concepts from several domains, including artificial intelligence, mainstream computing, mathematics, and human perception and cognition. The suggested SP system achieves lossless compression by matching and unifying patterns. However, this system is still in development (Belle et al. 2015).

Big data technologies for medical image processing

Parallel computing is detected as critical infrastructure for managing Big Data. It can perform analysis on a cluster of devices or supercomputers at the same time. Big Data technology with Artificial Intelligence (AI) and massively parallel computing can be used for a revolutionary way of prediction and personalized medicine (Dilsizian and Siegel 2014). Novel parallel computing models, such as Google's MapReduce (Dean and Ghemawat, MapReduce: simplified data processing on large clusters 2008), have been proposed in recent years for a new large data infrastructure. Apache just launched Hadoop (White 2015), an open-source MapReduce software for distributed data management. Concurrent data access to clustered servers is supported via the Hadoop Distributed File System (HDFS). Hadoop-based services may also be thought of as cloud computing platforms, allowing for centralized data storage as well as remote access through the Internet. As such, cloud computing is a revolutionary concept for distributing customizable computational resources across a network (Armbrust and Griffith 2010), and it may function as an infrastructure, platform, and/or software to provide an integrated solution. Furthermore, cloud computing may increase system speed, agility, and flexibility by eliminating the need to maintain hardware or software capacity and necessitating less resources for system maintenance, such as installation, setup, and testing. Cloud technologies are at the heart of many emerging Big Data applications (Luo et al. 2016). Additionally, Hadoop and Spark frameworks have been identified as optimal and efficient architecture for biomedical image analysis (Kouanou et al. 2018).

In addition, High Performance Computing (HPC) uses parallel processing and advanced programs, or software packages speed up massive calculations. In that sense, Finite Element Method (FEM), which represents continuum method for very powerful scientific computation analysis, strongly relies on advanced computer technology and HPC. Traditional database and software techniques cannot be used for these large-scale computations (Demchenko et al. 2013). High Performance Computing (HPC) can be used in medicine contained in Big Data (Lavignon et al. 2013).

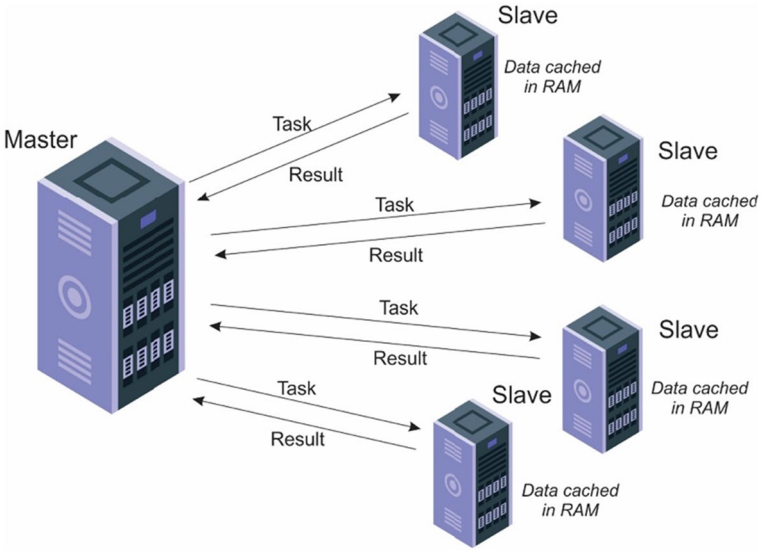

Fig. 2 Job execution using Spark technologies–one master cluster and four slaves

Massive multiscale computation with multiscale material models, or finite element computation with adaptive mesh refinement can be run only on supercomputers with Big Data on parallel disk systems (Parashar 2014). Detailed, complex, and anatomically accurate model of the whole heart electrical activity which requires extensive computation times, and the use of supercomputers are already established in the literature (Gibbons Kroeker et al. 2006; Kojic et al. 2019). The authors of this paper have recently developed a methodology for a real 3D heart model, by using the linear elastic and orthotropic material model based on Holzapfel experiments. Using this methodology, the transport of electrical signals and displacement field within heart tissue can be accurately predicted (Filipovic et al. 2022). Clinical validation in humans is very limited since simultaneous whole heart electrical distribution recordings are inaccessible for both practical and ethical reasons (Filipovic et al. 2022).

On the other hand, Apache Spark is a distributed computing platform that has become one of the most powerful frameworks in the Big Data situation. Spark provides a consistent and comprehensive framework for managing the needs for Big Data processing using a range of datasets (graph data, image/video data, text data, and so on) from various sources (batch, real-time streaming) (Tchito Tchapga et al. 2021). According to its designers, the Spark framework was intended to address the shortcomings of the Hadoop framework. In some cases, the Spark framework has shown to be quicker than Hadoop (more than 100 times in memory). Performance can be quicker than other Big Data technologies with advantages such as in-memory data storage and near real-time processing (Tchito Tchapga et al. 2021). The Spark framework can prepare data for iteration, query it frequently, and load it into memory. The main program (driver) in the Spark framework supervises many slaves (workers) and collects their results, whilst slaves' nodes read data partitions (blocks) from a distributed file system, run various computations, and write the results to disk (Fig. 2). This means that the master controls and assigns jobs to slaves.

Spark, like Hadoop, is built on parallel processing MapReduce, which seeks to process data in a simple and transparent manner across a cluster of computers. Spark enables SQL queries, streaming data, machine learning, and graph processing data in addition to Map and Reduce operations (Kouanou et al. 2018). In Spark, program can occasionally run the algorithm on several clusters at the same time. Although the number of slaves can be increased due to dataset size, the increase in the number of slaves results in an increase in processing time.

Existing solutions

Although there are studies that investigate Big Data in healthcare, to the best of our knowledge, no prior studies describe a complete procedure for managing medical images. Some of the existing research recognize Hadoop, as already mentioned, that uses MapReduce, as one of the frameworks built for analyzing and transforming Big Datasets (Shvachko, et al. 2010); Sobhy (2012). MapReduce is a programming paradigm that enables scalability across several servers in a Hadoop cluster for a wide range of real-world applications (Belle, et al. 2015); (Dean and Ghemawat, MapReduce: simplified data processing on large clusters 2008). However, it struggles with input-output heavy jobs (Markonis, et al. 2012) used the MapReduce framework to speed up three large-scale medical image processing use-cases: finding optimal parameters for lung texture classification using a well-known machine learning method, support vector machines (SVM), (ii) content-based medical image indexing, and (iii) wavelet analysis for solid texture classification. The entire execution time for obtaining optimal SVM parameters was lowered from around 1000 h to approximately 10 h (Markonis, et al. 2012).

Beside Hadoop, the other recognized technology is Spark. Kouanou et al. (2018) created a workflow that uses optimum algorithms integrating AI and ML to efficiently handle (acquire, analyze, process, distribute…) biomedical images. They present a comprehensive and effective process for managing biomedical images based on Big Data technology and optimal algorithms (AI and ML) derived from the literature. The classification phase in the suggested optimum flow will be treated as a study case utilizing Big Data analysis technologies (Hadoop and Spark) and may be adjusted to the remaining steps (Kouanou et al. 2018). They argue that Big Data applications frequently employ Not Only SQL (NoSQL) technologies (Sakr and Elgammal 2016; Bruchez 2015). NoSQL is a database category that debuted in 2009 and varies from relational databases (Bruchez 2015). One of the recurring issues with relational databases is the loss of performance while processing a high volume of data. Furthermore, distributed architectures necessitate native adaptation of solutions to data replication techniques and load control (Bruchez 2015; Lee et al. 2013). Cloud computing technology may also be utilized to help in data sharing because of the self-contained, networked IT (hardware and/or software) resources (Hassan 2011). When considering huge data, del Toro and Muller compared different organ segmentation approaches. They presented an approach that utilizes both the image's local contrast and atlas probabilistic information (del Toro and Müller 2014). When compared to using simply atlas information, an average of 33% improvement was obtained. Although some authors state that they investigate deep convolution neural network for Big Data medical image classification, the size of the datasets are really not Big Data, as number of images per class is 300 (12 classes total) (Ashraf et al. 2020a, b).

Other work is being done to manage and evaluate healthcare systems using Big Data. Behlima offered a Big Data management method for healthcare systems (Benhlima 2018). Kounau et al. (2018) proposed a new concept for biomedical image analysis using Big Data architecture in 2018. The authors present a workflow that performs the steps of acquisition of biomedical image data, analysis, storage, processing, querying, classification, and automatic diagnosis of biomedical images. The procedure used unstructured and structured image data from a NoSQL database. The authors developed a Spark architecture for constructing suitable and efficient techniques for classifying a huge number of photos. Belle et al. demonstrated the impact of Big Data analysis in healthcare (Belle et al. 2015); Luo et al. in (2016) conducted a literature review of Big Data application in biomedical research and healthcare; Viceconti et al. investigated the possibility of using Big Data for personalized healthcare (Viceconti et al. 2015); Archenaa and Anita (2015) demonstrated the need for Big Data analytics in healthcare to improve the quality of healthcare by providing patient-centric services and decentralized decision making. Thus, by incorporating Big Data technologies into a framework or applications, greater data handling and performance may be accomplished (Amanullah et al. 2020). Another publication by Tchapga et al. (2021) conducted a survey of biological image classification techniques. The paper then shows how to apply these techniques to a large data architecture using the Spark framework. They show that ML is vital in biomedical image classification, and when paired with Big Data technologies, the processing takes less time and can handle a large number of images at once. Although the advantages of Spark framework are big, the Spark framework's performance might suffer in some cases, most notably during feature extraction when there are some tiny images in the dataset (unlabeled biomedical images/labeled biomedical images). Another example is if the sizes of the images evaluated are too varied from one another, resulting in imbalanced loading in the Spark.

One of the articles by Cirillo et al. (2019) focuses primarily on analysis of multi-omics data as the main Big Data type in biomedical research and personalized medicine. Although these existing solutions each address one of the aspects of Big Data analysis (i.e. medical image processing), no solution has addressed the fully automated approach meaning the workflow starting from data (medical images), segmentation, 3D reconstruction, computational modelling and drug testing. To our knowledge, this paper is the first to present the complete workflow with details of methodology and results in each of the aforementioned fields.

Proposed solution for carotid artery ultrasound image processing

In this Section, we give a use case of ultrasound image processing, carried out by the authors of the paper during the TAXINOMISIS[1] project. We present the collected dataset during the project, proposed methods and obtained results. To this date, no full workflow of image processing and 3D reconstruction in a fully automated manner has been analyzed. The goal of TAXINOMISIS project is to create a new concept for carotid artery disease stratification by studying the pathobiology of symptomatic plaques,

[1] H2020 project TAXINOMISIS: A multidisciplinary approach for the stratification of patients with carotid artery disease, 755,320, 2018–2023, https://taxinomisis-project.eu/.

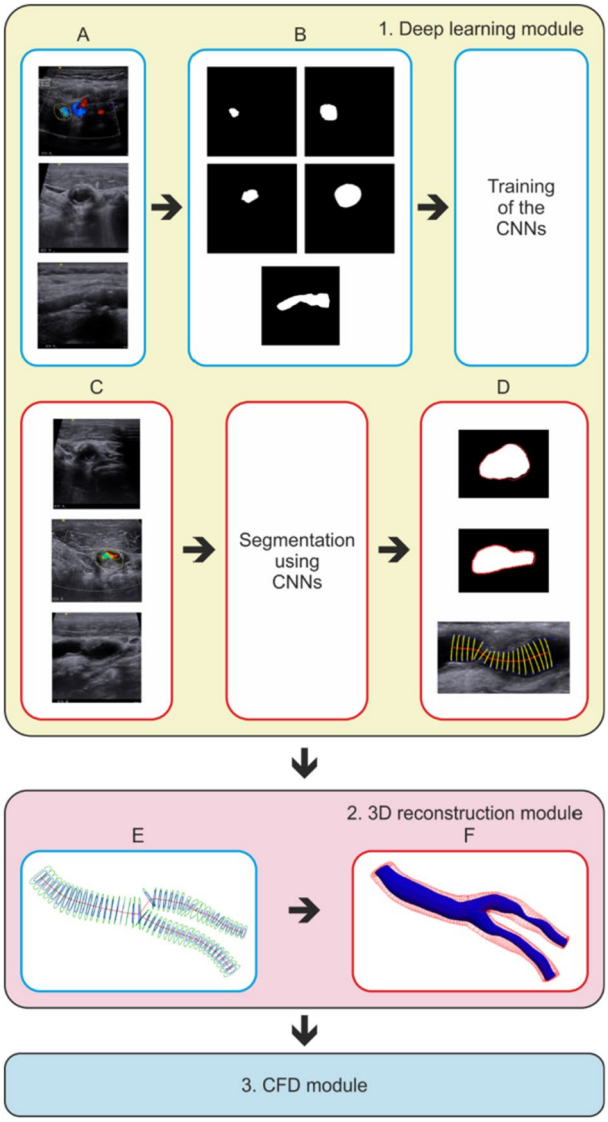

Fig. 3 Developed methodology for the reconstruction of carotid bifurcations using US imaging

identifying disease mechanisms, and developing a multiscale risk stratification model that integrates clinical and personalized data, plaque and cerebral image processing and computational modeling, and novel biomarkers for high vs. low risk states, in order to meet the need for stratified and personalized therapeutic interventions. Figure 3 depicts the overall pipeline used inside the project's US image processing module. The concept is collection of Big Data in terms of ultrasound (US) images of carotid artery with plaque (Fig. 3A). Collected image serve as the foundation for the whole reconstruction module. These images have been annotated and preprocessed (Fig. 3B). The convolutional neural networks (CNNs) are trained using these pairs of original and annotated images. The trained models are then utilized to extract the segments of the carotid artery (Fig. 3C). The deep learning module then directs the input to the reconstruction module (Fig. 3D), where the required forms of the carotid bifurcation are formed, as shown in Fig. 3E, in

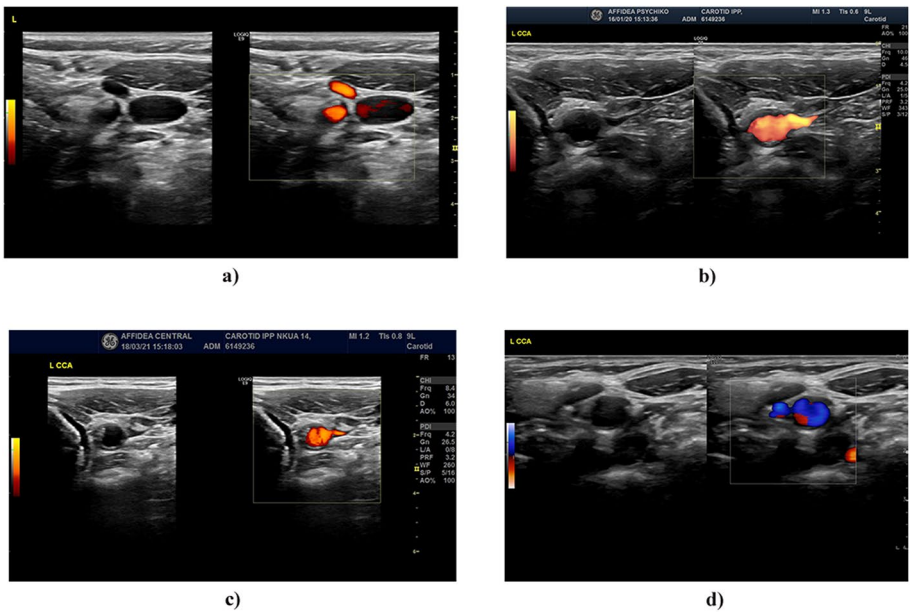

Fig. 4 Dataset samples

order to get the finite element mesh of the reconstructed geometry, as shown in Fig. 3F. Finally, the finite element mesh is ready to be employed in the computational fluid dynamics (CFD) module for blood flow and plaque progression simulations.

Dataset description

The dataset includes ultrasound images from 39 patients. At the carotid artery level, all subfolders corresponding to patients are randomly split into training, validation, and testing sets in an 8:1:1 ratio (either for the left or for the right arterial model). The dataset includes 179 original images that were supplemented for training purposes. When images from a completely new dataset (new slices received from another patient) are supplied as model input, random selection is critical to ensuring model robustness. This is both required and significant since the carotid arteries are not symmetric about the x and y axes. Furthermore, the patient's age, weight, and height are all crucial factors that determine the final outcome.

Image preprocessing

Given the low quality of US images, image preprocessing is a critical pre-step for many deep learning algorithms related to image and instance segmentation, as well as object recognition tasks. In this scenario, the automatic detection of the image region containing the artery tree under reconstruction is the initial step in preprocessing. This is accomplished by picking a 512×512 pixel static window for both artery models, left and right. The window coordinates are given special consideration so that the entire arterial tree is displayed in the region. Following this, all images are labeled, yielding two datasets with labeled areas, one for the lumen and the other for the wall.

Images in the dataset are extremely diversified as it is shown in Fig. 4. These differences include colors, brightness/contrast, frames and tables. It can be observed that some

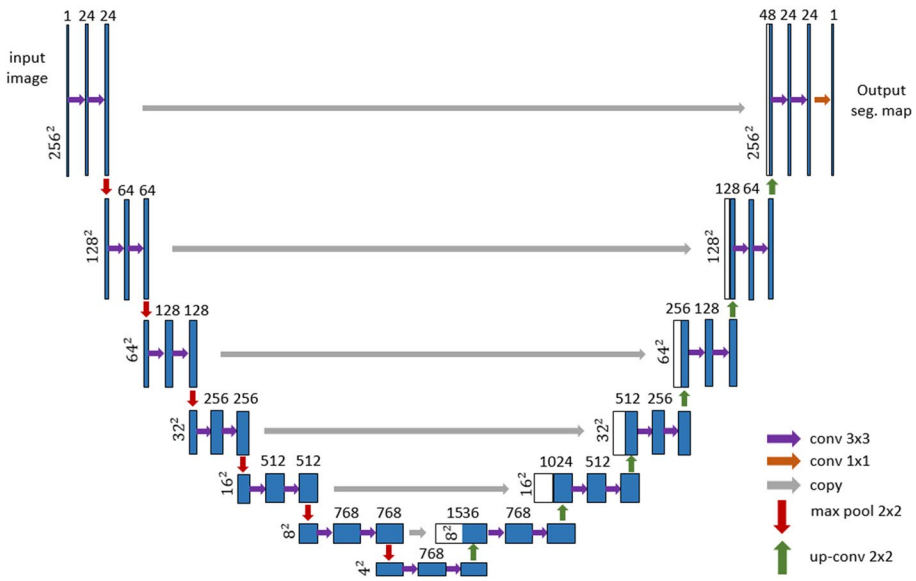

Fig. 5 U-Net architecture. Blue boxes represent the feature maps, white boxes are the feature maps copied from the encoder and concatenated with the decoder feature maps. Spatial resolution is shown on the left side of feature maps, and the number of channels is written on top of boxes

figures have tables containing various parameters on the right side (Fig. 4b and c), while others do not, and that color scales can be different (Fig. 4a and d). Additionally, as it can be seen Fig. 4d, significantly lower brightness/contrast of the figure caused overall dark appearance of the figure, where dark zones occupy a lot more space then on the other figures. Also, it can be observed that some figures have frames containing various information unlike others, and even that the spaciousness of the figures is different (Fig. 4b and c), i.e. they occupy less space on the figures.

Image segmentation

Image segmentation, more precisely automated carotid artery (lumen and wall) segmentation, was performed using FCN-8s, SegNet, and U-Net deep convolutional networks. In addition to the initial versions of these architectures, U-Net and SegNet networks were modified in terms of depth to assess their ability to distinguish regions of interest. Finally, the best results were obtained by employing a customized version of U-Net CNN (Fig. 5). It features two more blocks in both the encoder and decoder. Each encoder block comprises two convolutional layers with 33 filters, followed by 22 max pooling. Encoder blocks create output with 24, 64, 128, 256, 512, and 768 channels, respectively. In each decoder block, 22 upconvolution and skip connection are followed by three further convolutional layers with 33 filters, and the final decoder block creates the segmentation mask using 11 convolution and sigmoid activation function. All convolutional layers are padded such that the resultant activation map has the same height and width. As a result, the output segmentation map has the same resolution as the original image.

Furthermore, this version of architecture U-Net employs batch normalization after each convolutional layer, which performs significantly better on given data than the

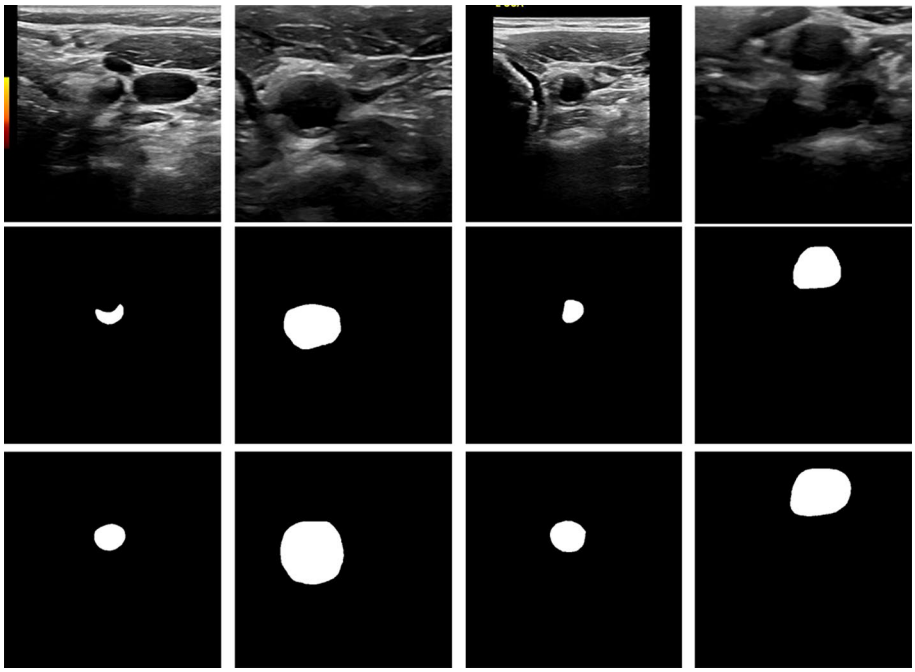

Fig. 6 Ultrasound imaging of the carotid arteries: original images (first row), lumen masks (second row), wall masks (third row)

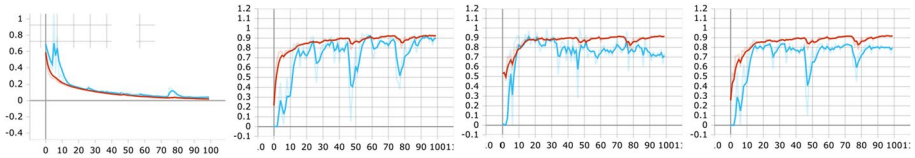

Fig. 7 The convergence of loss function, precision, recall and F1-score values over training (red) and validation data (blue)

Table 1 U-Net results on test dataset for lumen and wall

Target tissue	Precision	Recall	Dice coefficient (F1-score)
Lumen	0.89	0.78	0.83
Wall	0.82	0.91	0.85

original U-Net model. A ReLU activation follows each batch normalizing layer. As a loss function, the model is trained using a mix of binary cross-entropy and soft dice coefficient. In order to increase the number of images, to strive towards Big Data, data augmentation is performed. Because of relatively small size of the training set, the number of training photos was augmented further utilizing data augmentation techniques. Figure 6 shows examples of original and labeled images for the lumen and wall.

Figure 7 also depicts the convergence of the loss function for training and validation data for the lumen segmentation.

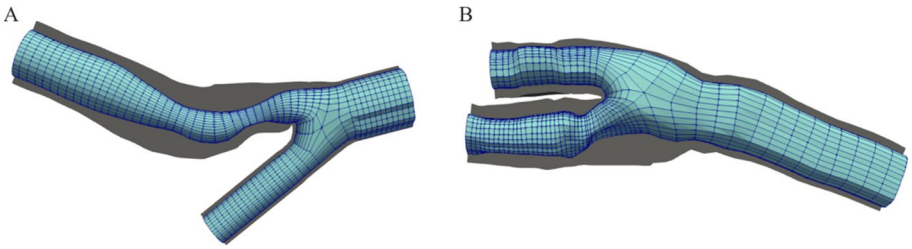

Fig. 8 Comparison of two reconstructed carotid bifurcations; **A** reconstructed geometry using with severe stenosis after bifurcation; **B** reconstructed geometry with mild stenosis after bifurcation

The results for binary classification were reported in terms of three standard metrics - precision (P), recall (R), and F1-score coefficient. Table 1 shows the findings of the lumen and wall for the test sets, respectively.

Training and testing were carried out on a GIGABYTE NVIDIA GeForce GTX 1080 Ti 11GB, GDDR5X, 352bit GPU. Python V3.6.7 is used as the programming language, while the code is written in Keras, which uses the Tensorflow framework as the backend.

3D reconstruction of carotid artery

The 3D reconstruction of the patient's carotid artery is accomplished utilizing the patient's available clinical imaging data. One of the most significant issues with the patient data set is the scarcity of 2D transversal slices. Furthermore, the longitudinal slices from the CCA and ECA branches were absent from the accessible dataset in the previously described approach. To address the issue of missing slices, the generalized model described in the literature was adopted as the foundation. This foundation is then tailored to the unique patient by incorporating accessible data into the geometry. Figure 8 shows the reconstructed model, where the elements of the model highlighted with a blue square have been adapted to the specific patient. The CCA and ICA branches' transversal cuts (annotated by the A, B, and D lines in Fig. 8B) are utilized to specify the forms of their cross-sections. The cross-section of the ECA branch is specified as circular since the clinical dataset lacked the transversal cut from this branch. The ICA's longitudinal cut is utilized to extract the centerline and diameters in this section, whereas the ECA and CCA branches are deemed straight. The arterial wall is also reconstructed using the available clinical data, in combination with generic data presented in literature, using the same approach that is used for lumen. Within the improved methodology, the longitudinal US images contained the whole carotid bifurcation. Hence the segmented data included lines of lumen and wall for all three branches. These lines are then used to define the shapes of all three branches where the parts of the model marked with a blue square are the ones that have been adapted to the particular patient. As it can be observed, the whole model is adapted to the particular patient. The lengths of the branches and their positions in space are also now patient-specific and not generic.

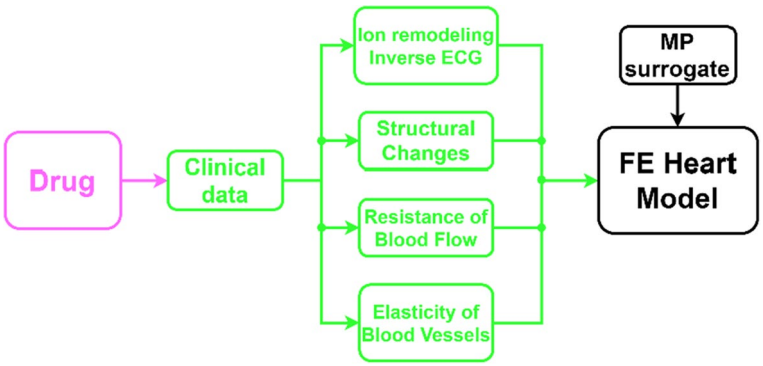

Fig. 9 Pathways of drug action in SILICOFCM drug testing workflow through macroscopic structural and boundary condition changes

Proposed solution for drug testing on personalized heart models

In this Section, we give a use case of computational modelling of the heart, carried out by the authors of the paper during the SILICOFCM[2] project. We present the collected dataset during the project, proposed methods and obtained results. To this date, no such complex methodology including computational modelling and drug testing has been performed. The challenge of the SILICOFCM project includes the development of a computational platform for in silico clinical trials of familiar cardiomyopathy that would analyze patient specific features (i.e. genetic, biological, pharmacologic, clinical, imaging and patient specific cellular aspects) with the aim to optimize medical treatment strategies. One of the workflows for testing of different types of drugs has been shown in Fig. 9. The experimental observations in many clinical trials are used as an input for FE models yielding us the precise model of Entresto®action.

Drugs that affect changes in macroscopic parameters-ENTRESTO®

ENTRESTO®(Sacubitril/valsartan) has been shown to be superior to enalapril in reducing the risks of death and hospitalization for heart failure (HF). There are also publications which evaluate the effects of sacubitril/valsartan on clinical, biochemical, and echocardiographic parameters in patients with heart failure and reduced ejection fraction (HFrEF). The first-in-class angiotensin receptor neprilysin inhibitor (ARNI) sacubitril/valsartan combines the angiotensin II type-1 receptor blocker (ARB) valsartan with the neprilysin inhibitor sacubitril. Entresto®was superior to enalapril in decreasing risks of death and new admission for HF in patients with HFrEF in the Prospective Comparison of ARNI with ACEI to Determine Impact on Global Mortality and Morbidity in Heart Failure (PARADIGM-HF) study (McMurray et al. 2014). Romano et al. (2019) investigated the effects of sacubitril/valsartan on clinical, biochemical and echocardiographic, parameters in HFrEF patients. They find that Entresto®can be "hemodynamic recovery" drug. A modulation of neurohormonal activation determined by this drug may lead to a hemodynamic effect that may impact cardiac hemodynamic and in association with Nt-proBNP concentration

[2] H2020 project SILICOFCM: In Silico trials for drug tracing the effects of sarcomeric protein mutations leading to familial cardiomyopathy, 777,204, 2018–2022, www.silicofcm.eu.

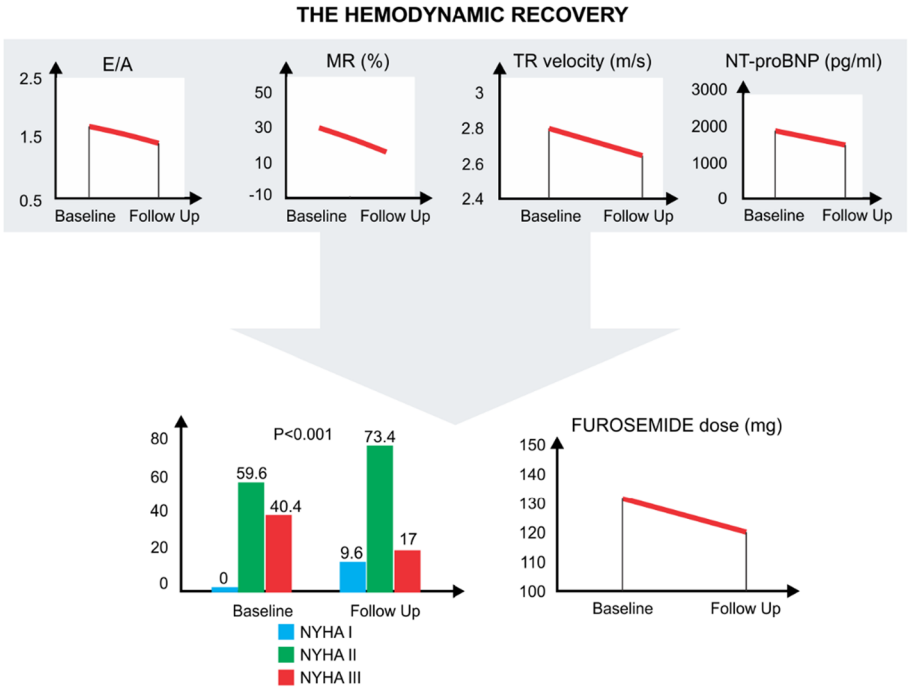

Fig. 10 Hemodynamic recovery

abatement could lead to a ameliorate NYHA (New York Heart Association) class and reduce diuretics administration and consequently to preserve renal function. Entresto®reduced E/A ratio, MR, TR velocity and Nt-ProBNP concentration. This hemodynamic effect ameliorates the NYHA class and reduce diuretic dose at follow-up. MR, mitral regurgitation from moderate to severe grade; E/A: peak e-wave velocity/peak a-wave velocity ratio; TR velocity: tricuspid regurgitation peak velocity (Fig. 10).

For example, the main Entresto®component, valsartan, strongly interacts with Angiotensin II receptor inducing the left ventricular hypertrophy in patients with essential hypertension. Left ventricular hypertrophy (LVH) represents an independent risk factor in patients with essential hypertension. In a randomized, double-blind trial, 69 predominantly previously untreated hypertensive patients with echocardiographically proven LVH, i.e., left ventricular mass index (LVMI) > 134 g/m^2 in men and > 110 g/m^2 in women and/or end-diastolic septal thickness > 12 mm, received the angiotensin II antagonist valsartan for 8 months (Thürmann et al. 1998). The study revealed that the dose of 80 mg/day decreased LVMI from 125 to 105 g/m^2 in 8 months. End-diastolic posterior and end-diastolic septal wall thickness also significantly decreased. Left ventricular end-diastolic and end-systolic diameter also decreased.

The influence of valsartan predominantly indicated significant increase in left ventricular end-diastolic and end-systolic volume, while ejection fraction remained in the same boundaries (Table 2).

Table 3 shows observed changes for Doppler echocardiographic parameters.

Pressure before and after Entresto treatment alongside the pressure volume diagram for left ventricle has been presented in Fig. 11.

Table 2 Influence of 80 mg/day dose of Entresto®

Parameter	0 months	8 months
LVMI—left ventricular mass index	125 g/m^2	105 g/m^2
PWTd—end-diastolic posterior wall thickness	13.6 ± 0.7 mm	12.4 ± 1.0 mm
IVSd—end-diastolic septal wall thickness	13.7 ± 1.2 mm	12.2 ± 1.1 mm
LVIDd—left ventricular end-diastolic diameter	47.24 ± 5.13 mm	46.22 ± 5.54 mm
LVIDs—left ventricular end-systolic diameter	29.07 ± 4.83 mm	28.46 ± 4.15 mm
FS—fractional shortening	39 ± 8%	38 ± 6%
LVEDV—left ventricular end-diastolic volume	91.00 ± 27.38 mL	94.97 ± 21.94 mL
LVESV—left ventricular end-systolic volume	31.31 ± 15.67 mL	34.07 ± 11.59 mL
EF—ejection fraction	65 ± 10%	65 ± 7%

Table 3 Doppler Echocardiographic Parameters

Parameter	0 months	8 months
$V_{max}E$, maximal velocity of early diastolic filling phase	75.30 ± 18.27 cm/s	70.14 ± 12.42 cm/s
$V_{max}A$, maximal velocity of late diastolic filling phase	82.64 ± 19.35 cm/s	78.07 ± 16.48 cm/s
$\int E$, time/velocity integral of early diastolic filling phase	10.76 ± 3.02 cm	11.61 ± 2.33 cm
$\int A$, time/velocity integral of late diastolic filling phase	10.66 ± 2.98 cm	10.46 ± 2.60 cm
$\int E/\int A$	1.06 ± 0.33	1.16 ± 0.29

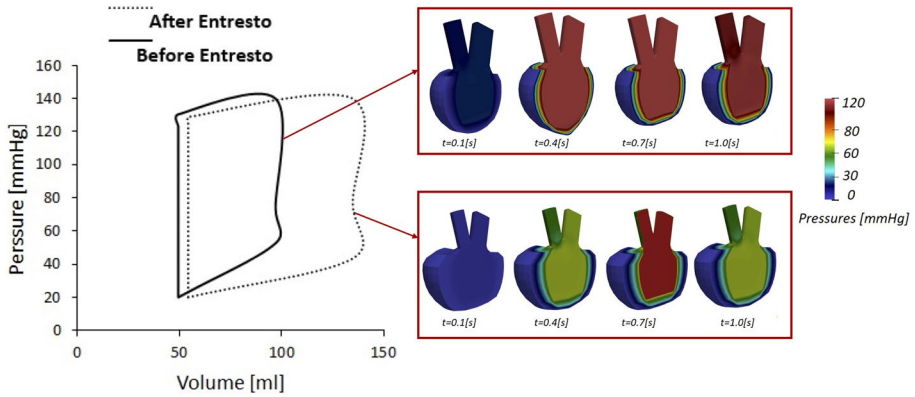

Fig. 11 Pressure volume diagram with details of pressure distribution before and after Entresto treatment

It is obvious that Entresto drug significantly increase ejection fraction for the left ventricle volume rate.

Realistic geometry of heart model with left chamber and atrium parts

Using experimental data and DICOM files provided from specific patient, we have reconstructed realistic heart model as STL format with left atrium (Fig. 12a, noted blue) and chamber part (Fig. 12a, noted yellow) with accompanying mitral valve cross-section between (Fig. 12a, noted green), and also aortic part (Fig. 12a, noted orange) of the model with aortic cross-section included in fluid part of the model, which is surrounded by solid wall (Fig. 12a, wireframe). Finite element model

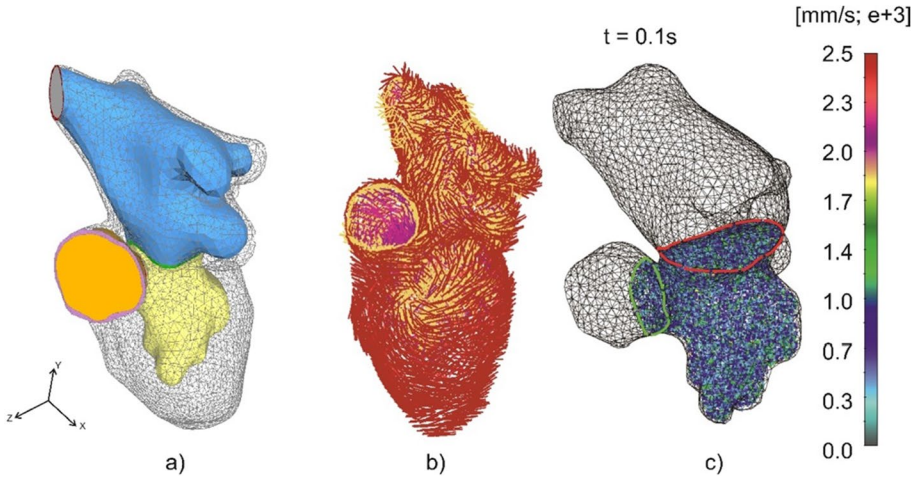

Fig. 12 **a** Realistic heart FE model with representative cross-sections and fluid parts; **b** Direction of fibres in solid part of realistic model; **c** Fluid velocity field at 0.1s (mitral and aortic cross-section noted)

consists of 1.5 M hexahedral 3D elements, divided by 1 M nodes. Model geometry is generated using STL files. Solid nodes are constrained around inlet/outlet cross-sections (Fig. 12a; red and magenta rings), and in the zone close to the mitral valve cross-section. Other solid nodes are free. In the Fig. 12c, two cross- section regions are marked to define prescribed inlet and outlet zones. Inside the fluid domain, mitral valve cross-section is presented (part of the model between ventricle and atrium; Fig. 12c, red line). Fibers direction in solid domain of realistic heart model are showed in Fig. 12b, and section C on the same figure shows distribution of velocity field in realistic heart model, at 0.1s. It can be seen that velocity values are the highest at inlet and outlet boundary cross-sections (red and green lines, Fig. 12c), which is logical due to prescribed inlet function and prescribed values at that cross-section at the beginning of simulation. Regarding the material models used, we have selected Holzapfel material model for obtaining passive stresses in the heart wall, and for muscle activation Hunter material model for active stresses is used.

Prescribed inlet velocity function profile is shown in Fig. 13a, and aortic valve cross-section, while outlet velocity function profile is shown in Fig. 13b. Activation of the muscle is achieved using calcium function, displayed in Fig. 13c.

Field of displacements in solid wall of realistic model of heart, during four different time steps of one cardiac cycle, is given in Fig. 14. At first step (0.1s), just the passive part of the material model has an impact on solid wall structure and until 0.4s of simulation model volume is increasing until the mitral cross-section is opened and fluid flows into the left chamber part. When the mitral valve is closed and injection of fluid is finished, fluid starts to eject from the chamber through the aortic cross-section, calcium function inside Hunter material model starts to act (0.5s), causing the start of the muscle contraction until the 0.9s of simulation after which model slowly returns to its undeformed state.

All of these models represent integration of Big Data technology, HPC and FEM computing.

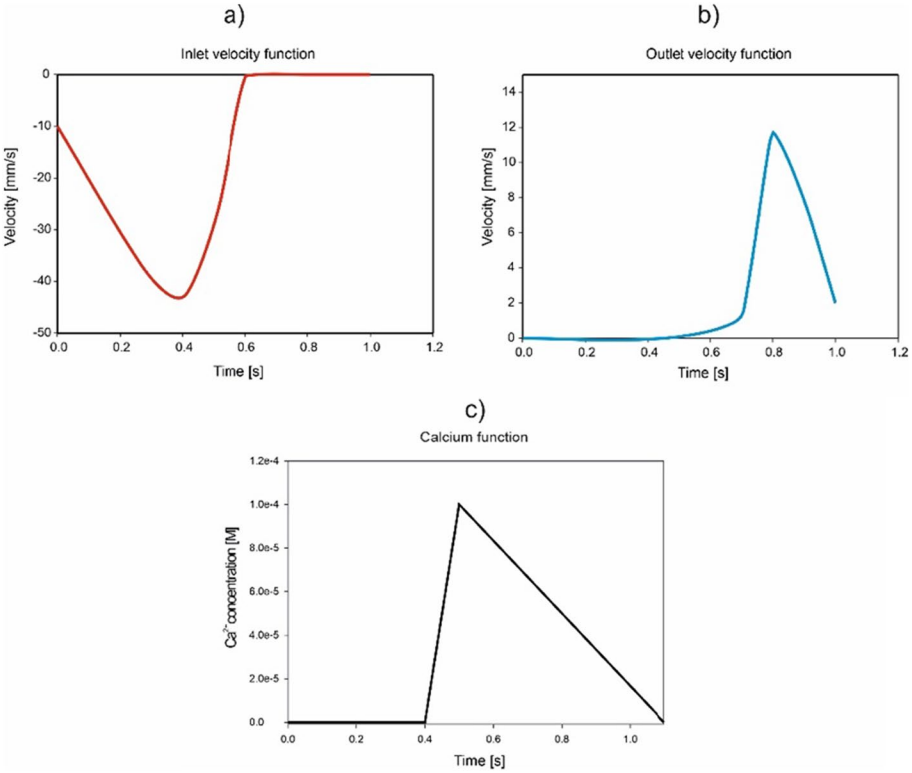

Fig. 13 **a** Inlet function of velocity, at mitral valve cross-section; **b** outlet velocity function–at aortic valve cross-section; **c** Calcium concentration function used for activation of the muscle

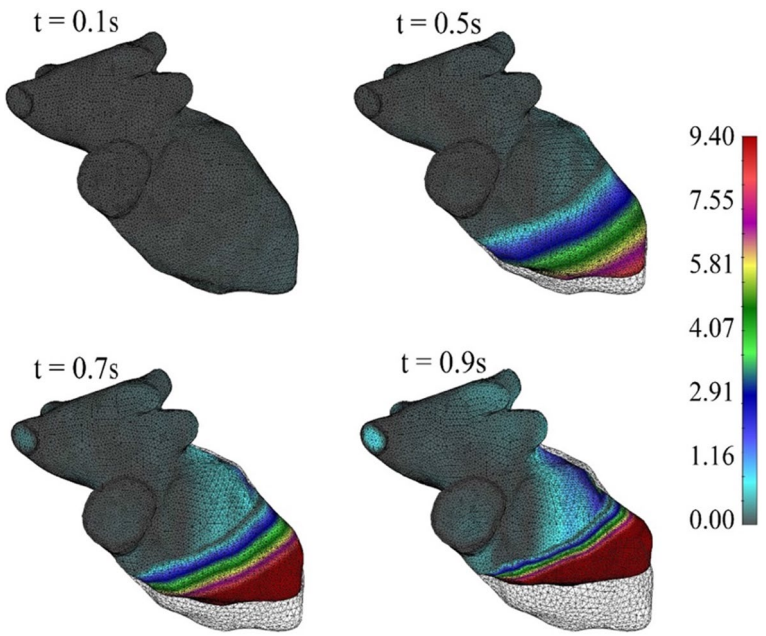

Fig. 14 Field of displacements in solid wall of realistic model of heart; four different time periods. Non-deformed configuration noted as black mesh

Conclusion

The availability of Big Data, novel deep learning algorithms and increasing computing power are three developments driving the deep learning revolution. While the potential advantages of deep learning are enormous, so are the early efforts and costs. Massive companies like Google DeepMind, IBM Watson, research laboratories, and prominent hospitals and vendors are collaborating to find the best solution for big medical imaging and image processing. Siemens, Philips, Hitachi, and GE Healthcare, among others, have already made substantial investments (Razzak et al. 2018). Similarly, research labs such as Google and IBM are investing in the delivery of efficient imaging applications, for example, IBM Watson is collaborating with more than 15 healthcare providers to discover how deep learning may be used in the real world (Razzak et al. 2018). Similarly, Google DeepMind Health is cooperating with the NHS in the United Kingdom to apply deep learning to various healthcare applications (for example, anonymized eye scans analysis might aid in the detection of disorders that lead to blindness) on a dataset of 1.6 million patients.

There are many papers discussing existing solutions related to Big Data and its application in medical domain, however, no studies have addressed complete workflow starting from data (medical images), image processing (i.e. segmentation), but also the aspect of 3D reconstruction, computational modelling and drug testing. To our knowledge, this paper is the first to present the complete workflow with details of methodology and results for two solutions (i) carotid artery ultrasound image processing and 3D reconstruction and (ii) *in-silico* drug testing on personalized models of heart. For both cases carotid artery and left ventricle model ultrasound image processing and 3D reconstruction was done using convolutional neural network U-net, while segmented masks were further used in 3D reconstruction of geometry. In the left ventricle and total heart personalized case computational finite element modelling was used to determine pressures before and after drug testing. Both cases demonstrate the necessity of using Big Data technologies.

There is no doubt that Big Data has great potential for improving health care. However, there are numerous challenges that healthcare faces when using Big Data technologies; the most major issue is the integration of multiple datasets. This gets more challenging when databases contain diverse data types (for example, integrating an image database or a laboratory test results database into current systems), restricting a system's capacity to query all databases to obtain all patient data. Future development of Big Data applications in healthcare domains offers a lot of potential since it is based on new data standards, appropriate research and technology, collaboration among research institutions and enterprises, and strong government incentives.

Acknowledgements
Not applicable.

Author contributions
Conceptualization, TG, DJ, NF; formal analysis, TG; funding acquisition, NF; investigation, TG; practical examples: TG, DJ, NF; visualization, TG and NF; writing—original draft, TG and NF; writing—review and editing, DJ, and NF. All authors read and approved the final manuscript.

Funding
The research was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, contract number [451-03-47/2023-01/200107 (Faculty of Engineering, University of Kragujevac)]. This research is also supported by the project that has received funding from the European Union's Horizon 2020 research and innovation programmes under Grant agreements No 755320 (TAXINOMISIS project) and No 952603 (SGABU project).

This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains.

Availability of data and materials
Data and materials for the cases running in this manuscript are deposited at https://taxinomisis-project.eu/ and www.silicofcm.eu (Accessed on 6 February 2023) for consortium members. The data are available on request from the corresponding author.

Declarations

Ethics approval and consent to participate
The use of imaging data as part of TAXINOMISIS project was approved by the Ethics Committee of the University of Belgrade (ClinicalTrials.gov Identifier NCT03495830 from September, 2019). The second part of the study related to SILICOFCM project was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the project (ClinicalTrials.gov, Identifier NCT03832660, from February 2019).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 February 2023 Accepted: 8 May 2023
Published online: 22 May 2023

References

Amanullah MA, Habeeb RA, Nasaruddin FH, Gani A, Ahmed E, Nainar AS,…, Imran M. Deep learning and big data technologies for IoT security. Comput Commun. 2020;151:495–517.
Andreu-Perez J, Poon C, Merrifield R, Wong S, Yang G. Big data for health. IEEE J biomedical health Inf. 2015;19(4):1193–208.
Archenaa J, Anita EM. A survey of big data analytics in healthcare and government. Procedia Comput Sci. 2015;50:408–13.
Armbrust M, A, F., Griffith Re. A view of cloud computing. Commun ACM. 2010;54(4):50–8.
Ashraf R, Habib MA, Akram M, Latif MA, Malik MS, Awais M,…, Abbas Z. Deep convolution neural network for big data medical image classification. IEEE Access. 2020a;8:105659–70.
Ashraf R, Habib M, Akram M, Latif M, Malik M, Awais M,…, Abbas Z. Deep convolution neural network for big data medical image classification. IEEE Access. 2020b;8:105659–70.
Belle A, Thiagarajan R, Soroushmehr S, Navidi F, Beard D, Najarian K. Big data analytics in healthcare. BioMed Res Int. 2015. https://doi.org/10.1155/2015/370194.
Benhlima L. Big data management for healthcare systems: architecture, requirements, and implementation. Advances Bioinform. 2018. https://doi.org/10.1155/2018/4059018.
Bruchez R. Les bases de données NoSQL et le Big data: comprendre et mettre en oeuvre. Editions Eyrolles; 2015. ISBN: 978-2-212-14155-9.
Cirillo D, Valencia A. Big data analytics for personalized medicine. Curr Opin Biotechnol. 2019;58:161–7.
Dean J, Ghemawat S. Map reduce: simplified data processing on large clusters. Commun ACM. 2008;51(1):107–13.
del Toro OA, Müller H. Multi atlas-based segmentation with data driven refinement. In. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). 2014; IEEE. (pp. 605–608).
Demchenko Y, Grosso P, De Laat C, Membrey P. Addressing big data issues in scientific data infrastructure. In 2013 International conference on collaboration technologies and systems (CTS). 2013; IEEE, New York City. (pp. 48–55).
Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep. 2014;16:1–8.
Doel T, Shakir DI, Pratt R, Aertsen M, Moggridge J, Bellon E, Ourselin S. GIFT-cloud: a data sharing and collaboration platform for medical imaging research. Comput Methods Programs Biomed. 2017;139:181–90.
Dutta H, Kamil A, Pooleery M, Sethumadhavan S, Demme J. Distributed storage of large-scale multidimensional electroencephalogram data using hadoop and hbase. Grid and Cloud Database Management. 2011; 331–47.
Filipovic N, Sustersic T, Milosevic M, Milicevic B, Simic V, Prodanovic M,…, Kojic M. SILICOFCM platform, multiscale modeling of left ventricle from echocardiographic images and drug influence for cardiomyopathy disease. Comput Methods Programs Biomed. 2022;227:107194.
George L. HBase: the definitive guide: random access to your planet-size data. O'Reilly Media, Sevastopol. Inc; 2011.
Gibbons Kroeker CA, Adeeb S, Tyberg JV, Shrive NG. A 2D FE model of the heart demonstrates the role of the pericardium in ventricular deformation. Am J Physiol Heart Circ Physiol. 2006;291(5):H2229–36.
Hassan QF. Demystifying cloud computing. Mansoura University, Mansoura. 2011.
Hersh W, Müller H, Kalpathy-Cramer J. The image CLEFmed medical image retrieval task test collection. J Digit Imaging. 2009;22:648–55.
Istephan S, Siadat M. Unstructured medical image query using big data–an epilepsy case study. J Biomed Inform. 2016;59:218–26.
Jai-Andaloussi S, Elabdouli A, Chaffai A, Madrane N, Sekkaki A. Medical content based image retrieval by using the hadoop framework. IEEE. 2013;2013:1–5.

Jayapandian CP, Chen CH, Bozorgi A, Lhatoo SD, Zhang GQ, Sahoo SS. Cloudwave: distributed processing of "Big Data" from electrophysiological recordings for epilepsy clinical research using Hadoop. AMIA Annual Symposium Proceedings. 2013; 2013, 691. American Medical Informatics Association
Jin Y, Deyu T, Yi ZA. Distributed storage model for EHR based on HBase. International conference on information management, innovation management and industrial engineering. IEEE. 2011; 2011(2): 369–72.
Kojic M, Milosevic M, Simic V, Milicevic B, Geroski V, Nizzero S. Smeared multiscale finite element models for mass transport and electrophysiology coupled to muscle mechanics. Front Bioeng Biotech. 2019;7:381.
Kouanou AT, Tchiotsop D, Kengne R, Zephirin DT, Armele NM, Tchinda R. An optimal big data workflow for biomedical image analysis. Inf Med Unlocked. 2018;11:68–74.
Lavignon JF, Lecomber D, Phillips I, Subirada F, Bodin F, Gonnord J, Muggeridge M. ETP4HPC strategic research agenda achieving HPC leadership in Europe. 2013.
Lee KK, Tang WC, Choi KS. Alternatives to relational database: comparison of NoSQL and XML approaches for clinical data storage. Comput Methods Programs Biomed. 2013;110(1):99–109.
Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomedical Inf insights. 2016;8:BII–S31559.
Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, Green ED. The National institutes of health's big data to knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014;21(6):957–8.
Markonis D, Schaer R, Eggel I, Müller H, Depeursinge A. Using MapReduce for large-scale medical image analysis. 2012 IEEE second international conference on healthcare informatics, imaging and systems biology. IEEE. 2012; 1–1.
Mazurek M. (2014, May 27–30). Applying NoSQL databases for operationalizing clinical data mining models. In. Beyond Databases, Architectures, and Structures: 10th International Conference, BDAS 2014. Proceedings 10 (pp. 527–536). Springer International Publishing, Ustron.
McMurray JJ, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR,. . ., Zile MR. Angiotensin–neprilysin inhibition versus enalapril in heart failure. N Engl J Med. 2014;371:993–1004.
Nguyen AV, Wynden R, Sun Y. HBase, MapReduce, and integrated data visualization for processing clinical signal data. AAAI spring symposium: computational physiology, palo alto. AAAI. California. 2011.
Ohno-Machado L, Bafna V, Boxwala AA, Chapman BE, Chapman WW, Chaudhuri K. iDASH: integrating data for analysis, anonymization, and sharing. J Am Med Inform Assoc. 2012;19(2):196–201.
Oussous A, Benjelloun F, Lahcen A, Belfkih S. Big data technologies: a survey. J King Saud University-Computer Inform Sci. 2018;30(4):431–48.
Pääkkönen P, Pakkala D. Reference architecture and classification of technologies, products and services for big data systems. Big data research. 2015;2(4):166–86.
Parashar M. Big data challenges in simulation-based science. DICT@ HPDC. 2014; 1–2.
Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: overview, challenges and the future. Classification in bio apps: automation of decision making. 2018; 323–350.
Romano G, Vitale G, Ajello L, Agnese V, Bellavia D, Caccamo G,. . ., Clemenza F. The effects of sacubitril/valsartan on clinical, biochemical and echocardiographic parameters in patients with heart failure with reduced ejection fraction: the "hemodynamic recovery. J Clin Med. 2019;8(12):2165.
Sahoo SS, Jayapandian C, Garg G, Kaffashi F, Chung S, Bozorgi A,. . ., Zhang GQ. Heart beats in the cloud: distributed analysis of electrophysiological 'Big data'using cloud computing for epilepsy clinical research. J Am Med Inform Assoc. 2014;21(2):263–71.
Sakr S, Elgammal A. Towards a comprehensive data analytics framework for smart healthcare services. Big Data Research. 2016;4:44–58.
Seibert JA. Modalities and data acquisition. Practical imaging informatics: foundations and applications for PACS professionals. 2010; 49–66.
Shvachko K, Kuang H, Radia S, Chansler R. The hadoop distributed file system. 2010 IEEE 26th symposium on mass storage systems and technologies (MSST). IEEE. 2010; 1–10.
Silva LA, Costa C, Oliveira JL. A PACS archive architecture supported on cloud services. Int J Comput Assist Radiol Surg. 2012;7:349–58.
Sobhy D, El-Sonbaty Y, Abou Elnasr M. MedCloud: healthcare cloud computing system. 2012 international conference for internet technology and secured transactions. IEEE. 2012; 161–166.
Tchito Tchapga C, Mih TA, Tchagna Kouanou A, Fonzin F, Fogang TKuetche, Mezatio P, B. A., Tchiotsop D. Biomedical image classification in a big data architecture using machine learning algorithms. J Healthc Eng. 2021. https://doi.org/10.1155/2021/9998819.
Thürmann PA, Kenedi P, Schmidt A, Harder S, Rietbrock N. Influence of the angiotensin II antagonist valsartan on left ventricular hypertrophy in patients with essential hypertension. Circulation. 1998;98(19):2037–42.
Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. IEEE J biomedical health Inf. 2015;19(4):1209–15.
Wang J, Qiu M, Guo B. Enabling real-time information service on telehealth system over cloud-based big data platform. J Syst Architect. 2017;72:69–79.
White T. Hadoop: the definitive guide. Sebastopol, CA: O'Reilly Media, Inc; 2015.
Wolfert S, Ge L, Verdouw C, Bogaardt MJ. Big data in smart farming–a review. Agric Syst. 2017;153:69–80.
Wolff JG. Big data and the SP theory of intelligence. IEEE Access. 2014;2:301–15.
Xing EP, Ho Q, Xie P, Wei D. Strategies and principles of distributed machine learning on big data. Engineering. 2016;2(2):179–95.
Yang A, Troup M, Ho JW. Scalability and validation of big data bioinformatics software. Comput Struct Biotechnol J. 2017;15:379–86.
Yao QA, Zheng H, Xu ZY, Wu Q, Li ZW, Lifen Y. Massive medical images retrieval system based on Hadoop. J Multimedia. 2014;9(2):216.

Zhang H, Wei X, Zou T, Li Z, Yang G. Agriculture Big Data: Research status, challenges and countermeasures. In Computer and Computing Technologies in Agriculture VIII: 8th IFIP WG 5.14 International Conference, CCTA 2014 Beijing. 2015
Zhang X, Yang Y, Shen L. Spark-SIFT: a spark-based large-scale image feature extract system. 2017 13th International conference on semantics, knowledge and grids (SKG) (pp. 69–76). IEEE. 2017.

Publisher's Note