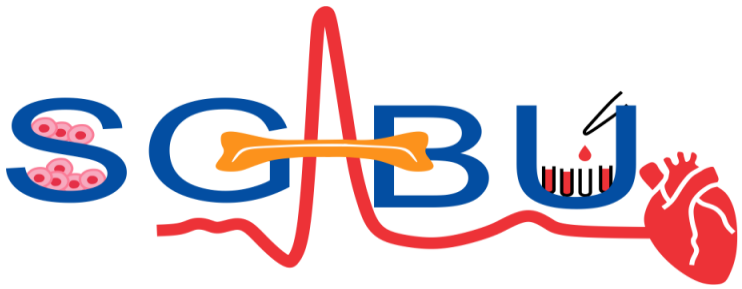

Project Title: *Increasing scientific, technological and innovation capacity of Serbia as a Widening country in the domain of multiscale modelling and medical informatics in biomedical engineering (SGABU)*

Coordinating Institution:
University of Kragujevac (UKG)

Start date: 1st October 2020

Duration: 36 months

Deliverable number and Title	**D3.2 State of the art in bioinformatics, imaging informatics, sensor informatics, public health informatics**
Related work package	**WP3**- Mobilize knowledge and expand the network
Related task	**Task 3.2** -State of the art in bioinformatics, imaging informatics, sensor informatics, public health informatics.
Lead beneficiary	UOI
Contributing beneficiaries	UOI
Deliverable type*	Report
Dissemination level**	Public
Document version	v1.0
Contractual Date of Delivery	31/01/2021
Actual Date of Delivery	28/02/2021

Authors	UOI (Orestis Gkaintes, Mairi Roumbi, Dimitrios Fotiadis)
Contributors	UOI
Reviewers	UKG (Nenad Filipovic, Aleksandra Vulovic), TUW (Christian Hellmich)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952603

Version history

Version	Description	Date of completion
0.1	General Document Structure and approach	12/12/2020
0.2	First draft consolidation	25/01/2021
0.3	Review	28/01/2021
1.0	Final version	28/02/2021

Disclaimer

*Deliverable Type: Report, Other, ORDP: Open Research Data Pilot, Ethics **Dissemination Level: PU=Public, CO=Confidential, only for members of the consortium (including Commission Services)

Executive summary

This deliverable provides all the information from the most current state-of-the-art methods in the field of machine learning, artificial intelligence, telemedicine, Internet of Things (IOT), bioinformatics, sensors, imaging to achieve patient specific outcome driven health care. More specifically, it includes technologies of devices, signals, and systems to optimize the acquisition, transmission, and storage for the biomedical and other related data. These analyzed methods will be part of the integrated SGABU platform, with focus on cardiovascular disease modelling and cancer modelling. The document focuses on issues such as how much information can be analyzed and retrieved from medical images (Imaging informatics), the developing methods and tools for a more comprehensive understanding of biological data (Bioinformatics), the wearable technologies for the prevention of diseases and the patient management (Sensor Informatics) and the systematic application of computer science and technology to public health practice (Public health informatics).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Explanation
ACP	American College of Physicians
ADNI	Alzheimer's Disease Neuroimaging Initiative.
AI	Artificial Intelligence
AI-CDSS	Artificial Intelligence-Clinical Decision Support System
AUC	Area under the receiver operating characteristic curve;
AODE	Averaged One-Dependence Estimators
ANOVA	Analysis of Variance (ANOVA)
BN	Bayesian Network
BBN	Bayesian Belief Network
BBT	Basal Body Temperature
BCG	Ballistocardiography
BFT	Byzantine Fault Tolerant
BM3D	Block matching into 3D data
CART	Classification and Regression Tree
CBF	Cerebral Blood Flow
CBR	Content-based retrieval
CC	Cloud Computing
CDISC	Clinical Data Interchange Standards Consortium
CDR	Clinical data repository
CCS	Circular Consensus Sequencing
CFD	Computational Fluid Dynamics
CGM	Continuous Glucose Monitoring
CHAID	Chi-squared Automatic Interaction Detection
ClinGen	Clinical Genomics

CNN	Convolutional Neural Network
CPs	Clinical Pathways
CPUs	Central processing units
CSER	Clinical Sequencing Evidence-Generating Research consortium
CM	Confocal Microscopy
CMBs	Cerebral Microbleeds
CLR	Continuous Long Read
CRT	Cyclic reversible termination
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CT	Computer tomography
DAM	Domain Analysis Model
DBM	Deep Boltzmann Machine
DBN	Deep Belief Networks
DCP	Dynamic Care Planning
DDD	Deciphering Developmental Disorders
DL	Deep Learning
DICOM	Digital Imaging and Communications in Medicine
DNA	Deoxyribonucleic acid
DNN	Deep Neural Network
DT	Decision Tree or Digital Twin (depending on the context)
DTI	Diffusion Tensor Imaging
dNTPs	Deoxynucleotides
dbGaP	Database of Genotypes and Phenotypes
dbSNP	Single Nucleotide Polymorphism Database
dsDNA	double-stranded DNA
ECG	Electrocardiogram
EEG	Electroencephalogram
EHRs	Electronic health records

eMERGE	Electronic Medical Records and Genomics (eMERGE) Network
EM	Expectation Maximization (algorithm)
EMR	Electronic Medical Records
EMBL-EBI	European Bioinformatics Institute
EMG	Electromyogram
ENCODE	Encyclopedia Of DNA Elements
EOG	Electro-oculogram
ETL	Extract, transform, and load
FCN	Fully Convolutional Network
FCRN	Fully Convolutional Residual Network
FDA	Food and Drug Administration
FDG	Fluorodeoxyglucose
FETs	Field effect transistors
FFT	fast Fourier transforms
FHIR	Fast Healthcare Interoperability Resources
FoG	Freezing of Gait
FT	Final Text
GA	*Genetic Algorithms*
GA4GH	Global Alliance for Global Health
GANs	Generative adversarial networks
GBM	Glioblastoma Multiforme
GBRT	Gradient Boosted Regression Trees
GDSC	Genomics of Drug Sensitivity in Cancer
GEO	Gene Expression Omnibus
GTEx	Genotype-Tissue Expression project
GWAS	Genome-Wide Association Studies
GIS	Geographic Information Systems
GMM	*Gaussian Mixture Model*
GPU	*Graphics Processing Unit*
GO	Gene Ontology
GSD	Ground State Depletion
GSR	galvanic skin response
gRNA	guide RNA
GWAS	Genome-Wide Association Studies
HIMSS	Health Information Management Systems Society
HL7	Health Level Seven

HMM	Hidden Markov Models
H&E	Hematoxylin and Eosin
HSI	Hyperspectral Imaging
IBSI	Image Biomarker Standardisation Initiative
ICA	Independent Component Analysis
ICD	Implantable Cardioverter Defibrillator
ICG	Impedance Cardiogram
ICGC	International Cancer Genome Consortium
IDFI	Incident Dark Field Imaging
ID3	Iterative Dichotomiser 3
IHE	Integrating the Healthcare Enterprise
IoT	Internet of Things
ILD	Interstitial lung disease
ISEs	Ion-Selective Electrode sensors
IT	Information Technology
IUT	Intersection-union test
kNN	k-Nearest Neighbour
LARS	Least-Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LED	Light-emitting diode
LDA	Linear Discriminant Analysis
LOESS	Locally Estimated Scatterplot Smoothing
LR	Logistic Regression
LRS	Long-read sequencing
LSTMs	Long Short-Term Memory Networks
LVQ	Learning Vector Quantization
LWL	Locally Weighted Learning
MARS	Multivariate Adaptive Regression Splines
MAS	Multi-Atlas Segmentation
MDA	Mixture Discriminant Analysis
MDR	medical records
MDS	Multidimensional Scaling
MESA	Multi-Ethnic Study of Atherosclerosis
MLP	Multilayer Perceptrons
MLR	Multiple Linear Regression
MRI	Magnetic Resonance Imaging
NB	Naïve Bayes
NCBI	National Center for Biotechnology Information
NNs	Neural Networks
NGS	Next generation sequencing
NLP	Natural language processing
NURBS	Non-uniform rational basis splines
NSIGHT	Newborn Sequencing in Genomic Medicine and Public Health program
OCT	Optical coherence tomography
OLSR	Ordinary Least Squares Regression

ONT	Oxford Nanopore Technologies
PAI	Photo-Acoustic Imaging
PALM	Photoactivation Localization Microscopy
PC	Public Comment
PCA	Principal Component Analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCI	Phase-contrast X-ray imaging
PCR	Principal Component Regression
PD	Parkinson's disease
PET	Positron Emission Tomography
PHR	Personal Health Record
PLSR	Partial Least Squares Regression
PNN	Probabilistic Neural Network
PoA	Proof of Activity
PoET	Proof of Elapsed Time
PoS	Proof of Stake
PoW	Proof of Work
PPG	Photoplethysmography
PSD	Power Spectral Density
PTT	Pulse Transit Time
QDA	Quadratic Discriminant Analysis
RA	Rheumatoid Arthritis
RBFN	Radial Basis Function Network
RESOLFT	Reversible Saturated Optical (Fluorescence) Transitions
RFs	Random Forests
RFID	Radio-Frequency Identification
RGB	Red-Green-Blue
RIP	Respiratory Inductance Plethysmography
RNA	Ribonucleic acid
RNNs	Recurrent Neural Networks
RS	Recommender Systems
SBL	Sequencing by ligation
SBS	sequencing by synthesis
SCG	Seismocardiography
SDF	Sidestream Dark Field
SDOs	Standard Developing Organizations
SIM	Structured Illumination Microscopy
SMR	Super-Resolution Microscopy
SNA	Single-nucleotide addition
SNPs	Single nucleotide poly-morphisms
SOM	Self-Organizing Map
SPECT	Single photon emission computed tomography
SPL	statistical pixel-level
SR	Synchrotron Radiation
SRA	Sequence Read Archive
SRS	Short-read sequencing

STED	Stimulated Emission Depletion
STORM	Stochastic Optical Reconstruction Microscopy
SVM	Support Vector Machines
TALENs	Transcription activator-like effector nucleases
TALN	Thoraco-abdominal lymph node
TBI	Translational Bioinformatics
TCGA	the Cancer Genome Atlas
TCIA	the Cancer Imaging Archive
TEG	Thermoelectric Generator
TF	Technical Framework
TI	Trial Implementation
TMA	Tissue Microarray
TILs	Tumor-infiltrating lymphocytes
TIRF	Total internal reflection fluorescence
TSP	Traveling Salesman Problem
UDN	Undiagnosed Diseases Network
UIT	Union-intersection test
US	Ultrasound
UV	Ultraviolet
VA	Virtual assistant
WCD	Wearable Cardioverter Defibrillator
WSI	Whole Slide Imaging
WNNM	weighted nuclear norm minimization
ZMW	Zero-mode wave guide
ZFNs	Zinc finger nucleases
4DCT	Four-dimensional computed tomography

1. Introduction

This deliverable collects all the state-of-the-art studies for enabling technologies of devices and sensors, databases, systems, signals, and big data analytics to refine the acquisition, processing monitoring, storage for biomedical and related social, behavior, environmental data.

Its aim is to present and analyze all the methods, mainly regarding the cardiovascular disease modelling and cancer modelling, which will be integrated into the SGABU platform.

According to the main document structure of the state-of-the-art deliverable the following sections are provided:

- Section 2 provides the state of the art in imaging informatics and specifically in Image Acquisition, Image Post Processing and Analysis in Radiology, Data Storage, Management and Sharing in Medical Imaging, Digital Pathology, In Silico Models and Integrative Analytics
- Section 3 provides the state of the art in bioinformatics and presents the current methods and platforms that are used for sequencing, as well as the stages that follow from the on stage of library preparation to data analysis and interpretation.
- Section 4 provides all the techniques and applications included in Sensor Informatics for healthcare problems.
- The last section describes current issues and proposed solutions in health domain, in the form of tools and software applications to support data collection, analysis and recording and advanced IT systems.

Finally, topics of privacy issues and challenges for health informatics are discussed.

2. The State of the Art in Imaging Informatics

Imaging informatics, also known as radiology informatics or medical imaging informatics, is a subspecialty of biomedical informatics that aims to improve the efficiency, accuracy, usability, and reliability of medical imaging services within the healthcare enterprise. It is devoted to the study on how information contained within medical images is retrieved, analysed, enhanced, and exchanged throughout the medical enterprise. In the following, an overview is provided of prevailing concepts, challenges and opportunities, and future trends are also discussed.

2.1 Image Acquisition

There are different image acquisition techniques reported in the literature, and with alterations in each modality regarding the technical setup, as well as the protocols to be followed. The most common modalities in imaging acquisition are the following:

Ultrasound (US): Acoustic waves at frequencies in the range of 1.5 - 15 MHz are transmitted into the body and the scattered and reflected echo-signals are processed to reconstruct an image. Several innovations extended the capabilities of ultrasound imaging: flow analysis by Doppler imaging led to the evaluation of the velocity profile in cardiovascular system [1]. To overcome the echogenicity of blood, ultrasound contrast agents made of encapsulated gas microbubbles were introduced in [2]. Attaching specific ligands to these bubbles enabled ultrasound molecular imaging [3]. Other applications of US include elastography, where the deformation of tissue is measured in the ultrasound images [4]. Ultrasound is not limited to 2D imaging, and use of 3D and 4D imaging is expanding, though with reduced temporal resolution [5]. A more recent innovation in contrast-enhanced imaging is ultrasound localization microscopy, where the localization of individual injected microbubbles and tracking of their displacements with a sub-wavelength resolution enables the production of vascular and velocity maps at the scale of µm [6] (Fig. 1). These techniques are now being applied pre-clinically and clinically for imaging of the microvasculature of the brain, kidney, skin, tumors, and lymph nodes.

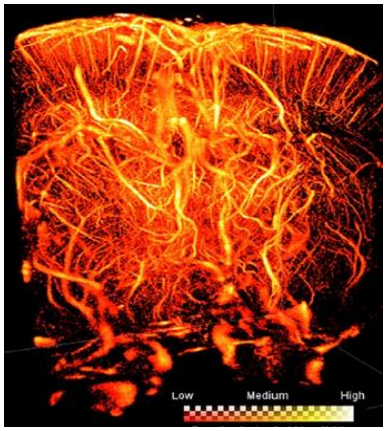

Figure 1: Volumetric ultrasound localization microscopy implemented on an anesthetized rat brain [6].

X-ray: Standard X-ray imaging techniques rely on a decrease of the X-ray beam's intensity when traversing the sample, which can be measured directly with the assistance of an X-ray detector. In Phase-contrast X-ray imaging (PCI) however, the beam's phase shift caused by the sample is not measured directly, but is transformed into variations in intensity, which then can be recorded by the detector [7]. To enhance the visibility of vascular structures and organs during radiographic procedure, radiocontrast agent containing iodine is granted to the subject in an intravenous form. X-ray projection imaging has been extensively used in cardiovascular, mammography and musculoskeletal imaging applications.

X-ray CT: X-ray computed tomography uses computer-processed combinations of multiple X-ray measurements taken from different angles to produce a 3D image via the construction of a set of 2D axial slices of the body. Modified versions of CT imaging, such as dual- and multi-energy CT, use additional attenuation measurements obtained with a second or multiple X-ray spectrum [8]. Four-dimensional computed

tomography (4DCT) is a type of CT scanning which records multiple images over time [9]. It allows playback of the scan as a video, so that physiological processes can be observed, and internal movement can be tracked. Recent developments within micro-computed tomography (micro-CT) imaging have combined to extend our capacity to image tissue in three (3D) and four (4D) dimensions at micron and submicron spatial resolutions, opening the way for virtual histology, live cell imaging, sub-cellular imaging and correlative microscopy [10]. CT scans have greater image resolution as compared to standard X-ray imaging, enabling examination of finer details, although there are concerns related to the required radiation dosage.

MRI: The operating principle of MRI is based on the detection of released electromagnetic energy, mainly from protons, due to the alteration of the magnetic field induced by the machine. Current versions make use of intravascular phase-contrast (PC) agents, and thus enable 3D velocity encoding, a technique known as 4D flow MRI [11]. 4D techniques are found in many applications for the visualisation of the cardiovascular system (Fig. 2).

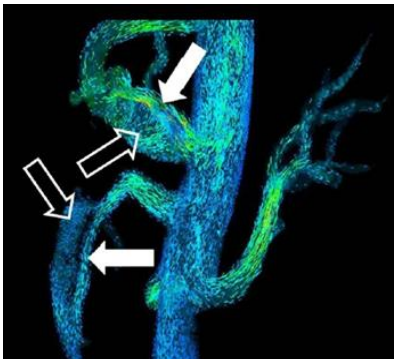

Figure 2: 4D PC MRI image depicting the fluid flow in a vascular region of a human [11].

Diffusion MRI: This technique utilizes the diffusion of molecules, mainly water, in tissues to generate contrast in images, in vivo and non-invasively [12]. A special kind of diffusion MRI, diffusion tensor imaging (DTI), has been used extensively to map white matter tractography in the brain [13].

Nuclear: Nuclear medicine uses radioactive tracers to assess bodily functions and to diagnose disease. Specially designed cameras allow doctors to track the path of these radioactive tracers. Single photon emission computed tomography (SPECT) and positron emission tomography (PET) are the two most common imaging modalities in nuclear medicine. Both techniques offer 3D image acquisitions generated by the computer from many projection (2D) images of the body recorded at different angles. The main difference between SPECT and PET scans is the type of radiotracers used. As far as their applications are concerned, SPECT scans are primarily used to diagnose and track the progression of heart disease, such as blocked coronary arteries [14]. There are also radiotracers to detect disorders in bone [15], gall bladder disease [16] and intestinal bleeding [17]. SPECT agents have recently become available for aiding in the diagnosis of Parkinson's disease in the brain [18] and distinguishing this malady from other anatomically related movement disorders and dementias. The major purpose of PET scans is to detect cancer and monitor its progression, response to treatment, and to detect metastases. Glucose utilization depends on the intensity of cellular and tissue activity, so it is greatly increased in rapidly dividing cancer cells. In fact, the degree of aggressiveness for most cancers is roughly paralleled by their rate of glucose utilization [19]. FDG has been shown to be the best available tracer for detecting cancer and its metastatic spread in the body [20].

Synchrotron radiation (SR): SR involves the tuning of ionizing beams of UV and X-ray over wide energy ranges. The interaction of the beams produced with molecules onto samples enables the production of detailed images of the molecular structure of the materials, as well as detailed chemical analysis [21]. SR has been applied in different areas of medical science, such as angiography, bronchography, mammography, microtomography (e.g., to study the 3D structure of trabecular bone), and in structural biology (e.g., protein crystallography), to name a few [22]. At the European Synchrotron Radiation Facility (Grenoble, France), a major research facility is operational on an advanced wiggler radiation beam port, ID17. The beam port is designed to carry out a broad range of research ranging from cell radiation biology to in vivo human studies [23].

Light Microscopy: The diffraction of light by small structural elements in a specimen is the principal process governing image formation in the light microscope [24]. Most light microscopes are limited by diffraction to about 250 nm resolution and magnifications by factor 2000. However, the more detailed architecture of unstained transparent biological samples cannot be seen with the normal bright-field microscope, which lead to the development of more intricate techniques such as fluorescence microscopy. Furthermore, two-photon fluorescence imaging uses two photons of similar frequencies to excite molecules which allows for deeper penetration of tissue. Progress in laser-scanning two-photon fluorescence microscopy permits fast 3D data acquisition [25], while several forms of volumetric microscopy are emerging, enabling the development of time-lapse studies (time-lapse microscopy) [26]. These technologies have been used in neuroscience [27], cancer [28], tissue imaging [29] and cell activity [30], among many other areas.

Confocal Microscopy (CM): CM is mainly used to produce images from the epidermis and superficial layer of dermis in human tissues [31]. In this technique, light is projected on a small area within the tissue, but unlike the conventional method the reflected light from the focal spot within the tissue is projected through a pinhole aperture onto a light detector. The use of a pinhole aperture allows only the focused light from the spot to pass through and eliminates the scattered light. Just as in conventional microscopy fluorescent dyes can be used to increase sensitivity and specificity. Confocal microscopy has been used to study the principles of vessel regression [32], erythrocyte properties [33], detecting skin lesions in oncology [34], and in studying the effect of glutaric acid on the blood–brain barrier and in perivascular astrocytes and pericytes [35]. It has also been used to quantify capillary cell blood flow [36] and in evaluating cutaneous microcirculation and dermal changes in systemic sclerosis [37].

Super-Resolution Microscopy (SMR): SMR is a family of techniques that emerged in the early 21st century. These techniques 'break' the diffraction limit that was previously thought to be impenetrable and as such allow for fluorescence imaging at resolutions up to ten times higher than in conventional techniques. Super-resolution imaging techniques rely on the near-field (photon-tunneling microscopy [38] and near-field scanning optical microscopy [39]) or on the far-field electromagnetic radiation physics. The major super-resolution microscopy techniques are stimulated emission depletion (STED), ground state depletion (GSD), reversible saturated optical (fluorescence) transitions (RESOLFT), photoactivation localization microscopy (PALM), stochastic optical reconstruction microscopy (STORM) and structured illumination microscopy (SIM). While SIM achieves a two-fold improvement in spatial resolution compared to conventional optical microscopy, STED, RESOLFT, PALM and STORM have all gone beyond, pushing the limits of optical image resolution to the nanometer scale ('nanoscopy'). Membrane nanostructure [40], dynamic regulation of proteins [41], chromosome dynamics [42], neuroscience (e.g., [43]), the visualisation of genome [44], and systems with relevance to the study of diseases associated with protein aggregation, including Alzheimer's [45] and Huntington's [46], are just a few examples where new insights have already been gained with super-resolution fluorescence nanoscopy.

Electron Microscopy: Electron microscopy uses a beam of accelerated electrons as a source of illumination. A scanning transmission electron microscope has achieved better than 50 pm resolution in annular dark-field imaging mode [47] and magnifications of up to about 10,000,000×. In combination with immunocytochemical methods, electron microscopy is a powerful method to label and highlight single specific proteins [48], enabling a correlation of the ultrastructure and its function.

OCT: Optical coherence tomography (OCT) is mainly used for cross-sectional tissue imaging [49]. This technique uses light in the near-infrared spectral range which has a penetration depth of several hundred microns in tissue. The backscattered light is measured with an interferometric set-up to reconstruct the

depth profile of the sample at the selected location. Advancements in OCT enable image flow [50], tissue dynamics [51] and even the estimation of mechanical properties like elasticity [52].

Photo-Acoustic Imaging (PAI): As the name implies, PAI is based on the photoacoustic effect [53]. This technique utilizes the detection of mechanical waves induced by the thermoelastic expansion of tissues, when subjected to non-ionizing laser pulses. To date, some clinical applications of PAI are found for label-free imaging of breast cancer [54], thyroid cancer [55] and inflammatory arthritis [56]. Instead of utilizing expensive lasers, low-cost light-emitting diodes (LED) are being investigated as a substitute laser illumination source in PAI systems [57]. AcousticX [58] is one such commercial LED-based PAI system.

A common characteristic between all the image modalities is the trade-off between the image resolution and the penetration depth. Figure 3 shows a diagram of resolution to penetration depth for selected image modalities. Of course, there are more image acquisition techniques, such as hyperspectral imaging (HSI), laser speckle contrast and laser Doppler perfusion imaging, side-stream dark field (SDF) and incident dark field imaging (IDFI), diffuse correlation spectroscopy, functional near-infrared spectroscopy etc.; a detailed description of which would make this study overwhelming.

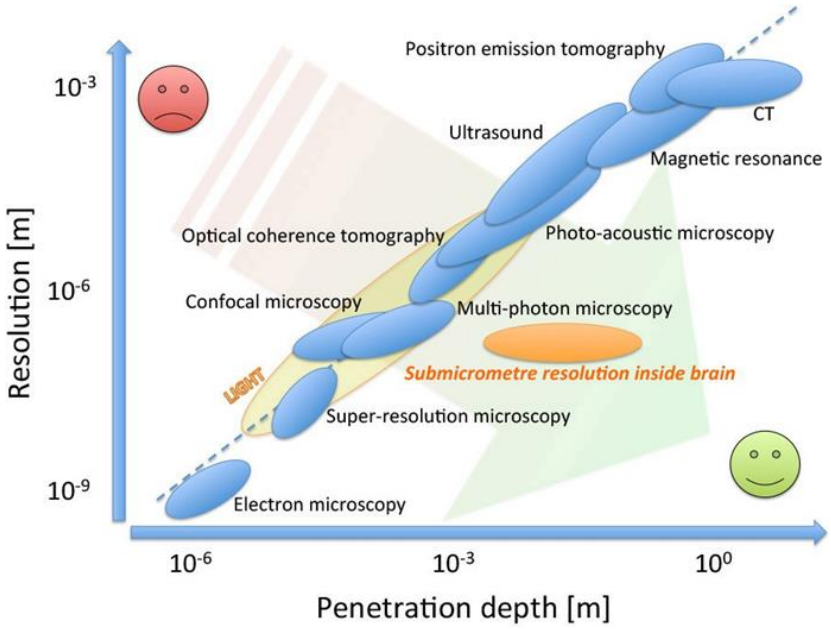

Figure 3. Diagram of resolution to penetration depth for different image modalities [59].

2.2 Image Post Processing and Analysis in Radiology

Medical image analysis involves the segmentation and classification of images. Image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyse. The intent of the classification process is to categorize all pixels in a digital image into one of several land cover classes, or ”themes”. This categorized data may then be used to produce thematic maps of the land cover present in an image. In the following, the current techniques applied in image segmentation and classification are presented.

2.2.1 Segmentation Process

The first step of processing images is to segment them. Lay Khoon Lee et al. [60] classify segmentation algorithms into four main types: 1) threshold techniques, 2) clustering techniques, 3) region and 4) edge localization models.

2.2.1.1 Algorithms Based on Thresholds

By assuming that an image is composed of multiple grey level regions and using a histogram to classify the different peaks and valleys of that image, thresholding-based segmentation techniques partition the pixels (of that image) depending on their intensity values. Algorithms of this type search for pixels whose values are within the ranges defined by selected either manually or automatically intensity-level thresholds. Manual selection needs a priori knowledge and sometimes trial experiments to find the proper threshold values while the automatic selection way combines the image information to get the adaptive threshold values automatically. The algorithms based on thresholds are the following:

(a) Local Thresholding:

Local thresholding determines different threshold values of sub-images by dividing an image into multiple sub-images or regions. Once each threshold is calculated, sub-images are merged. In addition, interpolation is applied to obtain appropriate results. The threshold value is calculated by different statistical methods, i.e. mean, standard deviation etc., applied on the histogram of the image to be segmented.

(b) Otsu's Method:

Otsu's method is used to perform automatic image thresholding [61]. The method works in determining an optimal value of threshold for segmenting the images. It is akin to calculating global threshold value, but Otsu's method takes into consideration inter-class and intra-class variation in an image.

(c) Gaussian Mixture Model (GMM):

GMM is not so much a model as it is a probability distribution. In general, GMMs are used for representing normally distributed subpopulations within an overall population. Mixture models do not require knowing which subpopulation a data point belongs to, allowing the model to learn the subpopulations automatically.

2.2.1.2 Region Based Segmentation

These algorithms are used to directly locate regions in an image based upon similarity [62]. They are split into two broad categories.

(a) Region Growing:

The family of these techniques involve in priori the selection of a seed pixel based on some characteristics (i.e. intensity level, inhomogeneity or edges in an image), and then the algorithm examines for neighbouring pixels that share a common characteristic property to determine whether the pixel neighbours should be added to grow the region. The process of growing a seed pixel is iterated until, e.g., an edge is detected.

(b) Region Splitting and Merging:

This technique first splits a given image into multiple sub-images and then merges them again. This approach is based on quad-tree generation, consisting of four branches. The branches of quad-tree represent sub-images [63]. The image region is split into four parts or branches and then merged back together till no partitioning or splitting is possible.

2.2.1.3 Edge/Boundary Based:

This method of segmentation deals with identifying and locating boundaries (or edges) in an image. The edges are sharp discontinuities (i.e. having different intensity values) in an image. The edge detectors are called "masks" or "filters" which are super-imposed over an image to detect discontinuities or boundaries [64]. The change in intensity level values of an image can be calculated by first order filters (Prewitt, Sobel, Canny) [65] that produce thick edges and second-order filters (Laplacian, Watershed Technique, etc.) that produce finer edges.

2.2.1.4 Clustering Methods:

Clustering methods are also considered as a sub-field of machine learning which is discussed below. This is a technique in which grouping of objects is done to form classes, and thus is referred to as clustering. The objects that share similar properties form a cluster. The objective here is maximisation of intra-class similarity and minimization of interclass similarity. It is a type of unsupervised learning as we do not need to train data. The widely used algorithms are the following:

(a) K-Means Algorithm:

K-means classifies the N datasets into k clusters iteratively. The mean intensity is calculated for each of the clusters and then the pixels are classified accordingly with closest mean values. This approach tries to reduce the number of clusters and cluster variability. It is usually used for segmentation of MRI images [66].

(b) Fuzzy C-Means algorithm:

This method is a generalisation of k-means algorithm and it is based on unsupervised learning. Since the uncertainty, vagueness and fuzziness are taken into consideration, it can sometimes result in introducing higher order fuzzy set for dealing with hesitation and uncertainty in classification of data.

(c) Expectation Maximization (EM) Algorithm:

EM algorithm is an iterative method to find (local) maximum likelihood estimates of parameters in statistical models (such as the GMM), where the model depends on unobserved latent variables. In each iteration, alternations between performing an expectation (E) step and a maximization (M) step are made to estimate the distribution of the latent variables.

2.2.1.5 Machine Learning:

Machine learning is a branch of AI which integrates data analysis methods to algorithms, used to perform a specific task without being explicitly programmed; Pattern recognition in the data and revision of predictions once new data arrive are the 'bread and butter' of AI. Usually, these methods need a large pool of training examples in order not to be subject to biases. Many algorithms of machine learning have been proposed in the field of imaging for segmentation and classification purposes. However, it seems until now that there is no such gold standard algorithm, and that the selection of the appropriate machine learning method depends on the application. In view of that, open access challenges such as the VISCERAL

project [67] are periodically organized in which participants can benchmark methods on standardized datasets. Table 1 gives a brief overview of machine learning algorithms.

Table 1. A summary of machine learning algorithms [1]

Group	Description	Selected algorithms
Regression Algorithms	Regression is concerned with modelling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model.	Ordinary Least Squares Regression (OLSR), Linear Regression, Logistic Regression, Stepwise Regression, Multivariate Adaptive Regression Splines (MARS), Locally Estimated Scatterplot Smoothing (LOESS)
Instance-based Algorithms	These models build up a database of example data and compare new data to the database to find the best match and make predictions.	k-Nearest Neighbor (kNN), Learning Vector Quantization (LVQ), Self-Organizing Map (SOM), Locally Weighted Learning (LWL), Support Vector Machines (SVM)
Regularization Algorithms	Regularizations work by making slight modifications to the learning algorithm so that the model generalizes better.	Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, Least-Angle Regression (LARS)
Decision Tree Algorithms	A model of tree-structure decisions is constructed, based on actual values of attributes in the data, until a prediction decision is made.	Classification and Regression Tree (CART), Iterative Dichotomiser 3 (ID3), C4.5 and C5.0, Chi-squared Automatic Interaction Detection (CHAID), Decision Stump, M5, Conditional Decision Trees
Bayesian Algorithms	They apply Bayes' theorem from statistics.	Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Averaged One-Dependence Estimators (AODE), Bayesian Belief Network (BBN), Bayesian Network (BN)
Association Rule Learning Algorithms	They try to extract rules that best explain relationships between variables in the data.	Apriori algorithm, Eclat algorithm
Clustering Algorithms	They use the inherent structure in the data to best organize the data into groups of maximum commonality.	k-Means, k-Medians, Expectation Maximisation (EM), Hierarchical Clustering
Dimensionality Reduction Algorithms	They are like clustering algorithms, but they work in an unsupervised manner, and 'describe' the data using less information.	Principal Component Analysis (PCA), Principal Component Regression (PCR), Partial Least Squares Regression (PLSR), Sammon Mapping, Multidimensional Scaling (MDS), Projection Pursuit, Linear Discriminant Analysis (LDA), Mixture Discriminant Analysis (MDA), Quadratic Discriminant Analysis (QDA), Flexible Discriminant Analysis
Ensemble Algorithms	They ensemble multiple weaker models that are independently trained and whose predictions contribute to the overall prediction.	Boosting, Bootstrapped Aggregation (Bagging), AdaBoost, Weighted Average (Blending), Stacked Generalization (Stacking), Gradient Boosting Machines (GBM), Gradient Boosted Regression Trees (GBRT), Random Forest
Artificial Neural Network Algorithms	They try to mimic the structure and function of biological neural networks.	Perceptron, Multilayer Perceptrons (MLP), Back-Propagation, Stochastic Gradient Descent, Hopfield Network, Radial Basis Function Network (RBFN)
Deep Learning	They work with deeper layers of neurons and they have a more complex form than conventional neural networks.	Convolutional Neural Network (CNN), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), Stacked Auto-Encoders, Deep Boltzmann Machine (DBM), Deep Belief Networks (DBN), CycleGAN

[1] The table is based on the book "Master Machine Learning Algorithms" by J. Brownlee, https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/.

2.2.1.6 Other Methods:

(a) Active Shape Models or Deformable Models:

In a deformable model approach, the shape of a model is optimized in order to match that of a structure of interest in an image [68]. This technique has been pioneered in 1987 by Terzopoulos et al. with the introduction of active contours or snakes [69]. This has been later generalized to active surfaces [70], but one difficulty arises when dealing with three-dimensional (3D) surfaces: the continuous parameterization of surfaces.

(b) Graph-Cut Method:

In this method, each image is represented as a graph of nodes, where each node corresponds to an image pixel and links connecting the nodes are called edges. After initialization of the path endpoints and adjustment of preferred paths by assignment of weights to individual edges, the algorithm tries to construct a pathway that minimizes the total weight sum [71].

(c) Level Set Method (LSM):

This technique is based on object and contour detection and curve evolution [72]. The advantage of the level-set model is that one can perform numerical computations involving curves and surfaces on a fixed Cartesian grid without having to parameterize these objects [73]. Also, the level-set method makes it very easy to follow shapes that change topology, e.g. when a shape splits in two or develops holes, making it a great tool for modelling time-varying objects.

(d) Multi-Atlas Segmentation (MAS):

MAS approach includes a wide array of sophisticated algorithms that employ ideas from machine learning, probabilistic modelling, optimization, and computer vision, among other fields. Atlases are labelled, manually by an expert, images intended to train algorithms for segmentation and classification purposes. In this respect, MAS offers several capabilities such as the flexibility to better capture anatomical variation, thus offering superior segmentation accuracy, although at high computational cost.

(e) Genetic Algorithms:

A genetic algorithm (GA) is a method for solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution [74]. The algorithm works in 3 steps referred as operators: mutation, crossover, and selection. The method starts from a group of solutions (initial population), then evaluates the fitness of each individual in the population and repeats on selection of best until termination. The best individuals from the population are then combined to produce the offspring which possesses better characteristics. The changes, or mutations, introduced result in generation of heuristic solutions from the population until the most optimised solution is obtained. In image processing GAs have been used for image enhancement, segmentation, and feature extraction.

2.2.2 Classification Process

Classification is the process of finding patterns on image features from numerical databases and categorizing pixels in an image into classes by providing suitable class labels. According to [75] the classification process typically involves the following steps: image pre-processing, feature extraction, selection and classification (Fig. 4).

Pre-processing is the first step in the classification process and its aim is to suppress unwanted distortions and enhance image features important for further processing. Commonly applied steps in this stage include: 1) the conversion of the RGB image into a grayscale image, or 2) into a two-color pixel (usually black and white) image, 3) the improvement of contrast in the image (contrast stretching), 4) noise removal of unwanted artifacts, and 5) sharpening of the image (i.e. creating an image that is less blurry than the original).

Feature extraction involves the reduction of features extracted from the image. The techniques in feature extraction are usually based on 1) statistical pixel-level (SPL) features (such as mean, variance etc.), 2) the color histogram of the pixels, 3) shape features (e.g. features providing information about the characteristic of the region boundary), 4) texture features that characterize the spatial distribution of intensity levels in the local region of interest, and 5) relational features that provide information about the structure of the image with respect to single or multiple objects.

The last step prior to classification is the selection of important features that are deemed important for the classification task. Different methods in feature selection include 1) filter methods, such as Person's correlation, linear discriminant analysis (LDA), analysis of variance (ANOVA), principal component analysis (PCA) and chi-square test, 2) wrapper methods that are based on machine learning approaches, such as forward selection, backward elimination and recursive feature elimination, and 3) embedded methods which combine wrapper and filter methods, such as least absolute shrinkage and selection operator (LASSO).

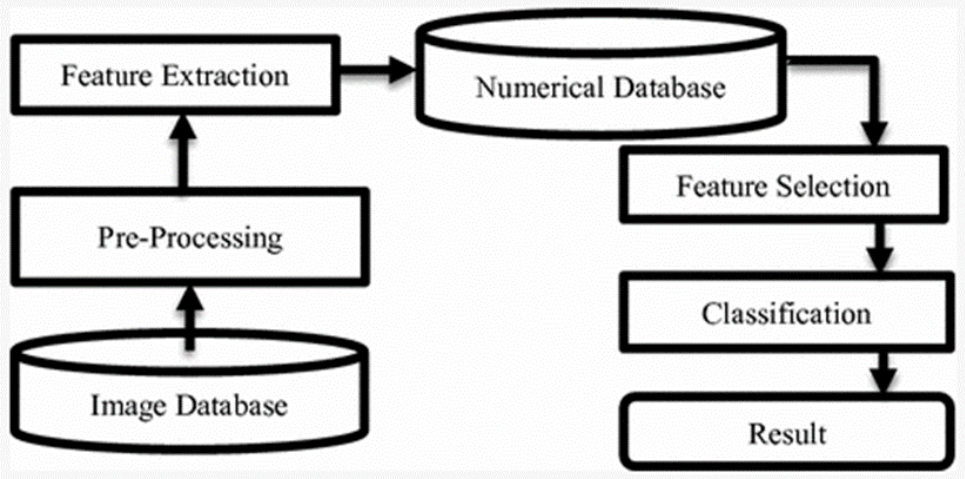

Figure 4: A generic overview of classification process [75].

2.2.2.1 Classification Techniques

In the following, table 2 summarizes widely used classification techniques along with their benefits and limitations.

Table 1: A summary of widely used classification algorithms**.**

Classification techniques	Description	Benefits	Limitations
Decision tree induction [184]	The dataset is broken down into smaller subsets and is present in the form of tree nodes. The tree structure is characterized by a root node, decision nodes, leaf nodes and branches. ID3 and C4.5 are decision tree algorithms.	○ It requires less effort for data preparation during pre-processing. ○ It does not require normalization and scaling of the data. ○ Missing values in the data do not considerably affect the process of building the decision tree ○ Learning and classification steps of a DT are simple and easy to explain.	• Instability, i.e. a small change in the data can cause a large change in the tree structure. • It needs big datasets and more time to train the model. • As the dataset grows larger, tree calculations get complex, thus making it not much useful in practical approaches. • As the dataset gets growing tree calculations get complex, thus making it not much useful in practical approaches. • Inadequate for applying regression and predicting continuous values.
Bayes classification methods [185]	They are statistical classifiers based on Baye's Theorem. They calculate class membership probabilities.	○ Easy and fast to implement. ○ It scales linearly with the data size, making it capable of handling large datasets. ○ Small memory footprint. ○ Noise Resilience, i.e. if the data has noise or irrelevant features its capabilities will not be seriously affected. This implies also that there will be less risk for overfitting. ○ It can train with a small dataset. ○ It can be updated ``on the go'' and quickly. ○ It provides useful outputs such as mean and variance values for each feature and class.	• It assumes that the features are independent. In real world problems where the correlation between the features is high, bias and inaccuracy will appear. • Inadequate for applying regression. • Training with more data will not make the algorithm capable of making complex predictions.
Rule-based classification [190]	It contains set of IF-THEN rules. The IF part involves one or more attribute tests, and these tests are logically ended, and ELSE part involves class prediction. One rule is	○ Training data not required. ○ It comes handy to collect data as one starts the system with rules. ○ High precision.	• The modification of knowledge base can be complicated. • High computational cost. • Very difficult to examine what actions are going to happen, and when.

	created for each path ranging from the root to the leaf node.		• It can result to complex domains.
Neural network classifiers [186]	It consists of units (neurons), arranged in layers, which convert an input vector into some output, trying to mimic neural networks of biological systems. In general, neural networks are defined as feed forward, back propagation, radial basis function, recurrent neural network, etc.	○ It performs well for nonlinear data with large number of inputs. ○ Once trained, the predictions are fast. ○ It can handle complex datasets, and even if the dataset is augmented then it will still provide finer results.	• High computational cost. • Time consuming with traditional CPUs. • It depends on large training datasets, making the training process slow. Initial tuning may be needed. • It is prone to overfitting.
Support vector machines (SVM) [187]	A support vector machine is a binary classifier, which uses kernel function to transform low-dimensional training samples to higher, and quadratic programming to find the best classifier boundary hyper-plane. It can incorporate expert knowledge by using the kernel trick.	○ More effective in high dimensional spaces. ○ High accuracy in classification. ○ Works well when there is a clear margin of separation between classes. ○ It maximizes margin, so the model is more robust. ○ It supports kernels, enabling model design for even nonlinear relations. ○ Fewer parameters to consider (kernel, error cost C) compared to neural networks. ○ Works well with fewer training samples.	• Hard to interpret. • Not suitable for large data sets. • It does not perform well when target classes overlap, and when the number of features for each data point exceeds the number of training data samples. • Memory intensive.
Lazy learners [188]	It refers to machine learning methods in which generalization of the training data is delayed until a query is made to the system. The outcome of these algorithms is always class label. K-nearest is a good example of these algorithms.	○ Simple and easy to implement. ○ Parameters are not required. ○ High accuracy, but lower compared to other supervised learning models.	• High memory requirement. • High computational cost. • If dataset is nonlinear, then it is not working. • Hard to find optimal value, still not guaranteed about optimal solution.

Deep Learning [189]	It is like conventional neural networks but consists of more layers of neurons and presents more complexity. Convolutional neural network (CNN) belongs to DL methods.	○ Features are automatically deduced and optimally tuned for desired outcome. ○ Features are not required to be extracted ahead of time. ○ It offers robustness. ○ Massive parallel computations can be performed using GPUs and are scalable for large volumes of data. ○ It delivers better performance results when amount of data is huge. Flexible architecture to be adapted for different scenarios.	• It requires very large amount of data to perform better than other techniques. • It is extremely expensive to train due to complex data models. It also requires expensive GPUs and hundreds of machines. • They are black-box functions and their output is difficult to comprehend. • There is no standard theory to guide you in selecting right deep learning tools as it requires knowledge of topology, training method and other parameters, making it difficult to be adopted by less skilled people.

2.2.3 Deep Learning for Segmentation/Classification

Among the machine learning algorithms, deep learning seems to represent the current state of the art in image processing. The introduction of AlexNet [76] offered dense predictions obtained from classification networks and set the standard for image classification. With that, other deep learning methods followed, such as Deep Boltzmann Machines and stacked autoencoders, offering revolutionary results in segmentation of anatomy and pathology. Another notable example is CycleGAN which treats image-to-image segmentation as a translation and synthesis problem, and it allows mapping of one image domain to another image domain even without having pairs of images [77]. Furthermore, CNNs have been used for registration tasks in medical image analysis. In their work, de Vos et al. [78] propose an entire framework for unsupervised affine and deformable image registration, in which once learning is done, CNNs can register unseen images in one shot.

However, most deep models have millions to billions of parameters, needing a vast amount of data to optimize them. The discovery of 2D U-Net addresses this issue [79], offering highly accurate segmentation boundaries even with few input data, while recently the development of 3D U-net was proposed that allowed full volumetric processing of imaging data [80]. Other methods include fine-tuning a pre-trained model, which speeds the training process. However, this transfer learning approach is not straightforward when the objective is tissue classification of 3D image data. Here, transfer learning from natural images is not possible without first condensing the 3D data into two dimensions [81]. However, some approaches directly exploit the 3D data by using architectures that perform 3D convolutions and then train the network from scratch on 3D medical images [82]. Table 3 presents applications of DL methods for segmentation/classification purposes in different scenarios of medical imaging.

DL methods, such as CNN algorithms, are black-box functions and it is not easy to comprehend their output. What the CNN network has learned and how it derives its classification decisions is an emerging area of deep learning. Common approaches include 1) the visualization of nearest neighbors of image patches [76], 2) the creation of saliency maps [83], guided propagation [84], and 3) the feature inversion approach [85]. Lastly, worth to mention, MICCAI [86], CAMMELYON [87] and IEEE [88] challenges are posted every year for segmentation and classification algorithms for images obtained with different modalities.

Table 2: Different DL methods applied for classification/segmentation purposes [89].

Year – [REF] Author	Disease	Imaging Data	DL method	Segmentation/ Classification	Description
2015 – [79] Ronneberger et al.	Cells	Electron and optical microscopy	U-net	Segmentation of images and cell tracking	The proposed network and training strategy can classify objects based on few annotated samples. Ranked 1st place in the IBSI cell tracking challenge 2015.
2016 – [90] Shin et al.	LN, ILD	CT	Transfer Learning (AlexNet, GoogleNet, CifarNet CNNs)	Thoraco-abdominal lymph node (LN) detection and interstitial lung disease (ILD) classification	The authors studied different CNN architectures (i.e., they varied the number of parameters and layers), the influence of dataset scale and spatial image context, and whether transfer learning from pre-trained ImageNet could be useful.
2016 – [91] Dou et al.	Cerebral Microbleeds (CMBs)	MRI	Two-stage: 1) 3D Fully-Convolutional network (FCN), 2) 3D CNN	CMBs detection	To reduce computational costs, the proposed framework used a 3D FCN to detect high probability candidates of CMBs, and then a trained 3D CNN to distinguish CMBs from mimics. 93.16% sensitivity and a mean number of 2.74 false positives per subject were achieved.
2016 – [81] Setio et al.	Pulmonary Cancer	CT	Two-stage: 1) Feature-engineered candidate detector, 2) Multi-view 2D CNN for false positive reduction	Candidate pulmonary nodules detection	Volumetric images are firstly decomposed into fixed triplanar views (sagittal, coronal, and axial planes), then each plane is processed with 2D CNNs, and their output is combined to make the final classification. The method reached nodule-detection sensitivity 85-90% and 1-4 false positive per scan.
2017 – [92] Lekadir et al.	Cardiovascular (carotid artery)	US	Four convolutional and three fully connected layers	Characterization of carotid plaque composition	An automated technique, using CNNs, was proposed to discriminate between different plaque constituents. Cross validation results showed 90% correlation with the clinical assessment.
2017 – [93] Yu et al.	Melanoma	Dermoscopic Images	Deep (38/50/101 layers) fully convolutional residual network (FCRN)	Binary melanoma segmentation and classification	First, residual learning is applied to cope with degradation and overfitting problems. Then, a FCRN is constructed, integrating contextual information for skin lesion segmentation. Finally, the FCRN is integrated with other deep residual networks to form a two-stage framework for classification. The proposed framework ranked 1st in classification and

					2nd in segmentation, in IBSI challenge 2016 for skin lesion analysis towards melanoma detection.
2017 – [94] Lao et al.	Glioblastoma Multiforme (GBM)	MRI	CNN and Transfer Learning	Segmentation of three tumor sub-regions, including the necrosis area, the enhancement area and the edema area	In this framework, both hand-crafted features and deep features were extracted from multi-modality MR images, to construct a radiomic signature for prediction of overall survival in patients with GBM. Deep features were extracted from pre-trained CNN via transfer learning. Combining the signature with other risk factors, the combined model achieved about 70% predictive performance.
2017 – [95] Oakden-Rayner et al.	Overall Survival	CT	CNN transfer learning (3 convolutional and 1 fully connected layers)	Tissue (muscle, body fat, aorta, vertebral column, epicardial fat, heart, lungs)	This study demonstrates how CT images combined with computer-aided systems can be used to predict longevity.
2017 – [96] Zhu et al.	Breast cancer	DCE-MRI	Transfer learning (GoogleNet, VGG-Net, CIFAR)	Breast tumor lesions	Three different DL approaches - training from scratch, transfer learning and off-the-shelf deep features – were applied to discriminate between different breast cancer subtypes. The results were validated using 10-fold cross-validation and area under the receiver operating characteristic (AUC). Off-the-shelf deep features approach achieved the best AUC performance of 0.65 (95% CI: [0.57,0.71]).
2018 – [97] Chartsias et al.	Cardiovascular	MRI	Various Networks	Segmentation of cardiac anatomy	The authors propose a method for disentangling medical images, entailing anatomical information and properties related to imaging setting. Furthermore, they demonstrate the efficacy of their method in a semi-supervised myocardium segmentation task, achieving comparable performance to fully supervised networks, by using a fraction of labelled images for training.
2020 – [98] McKinney et al.	Breast Cancer	X-ray	Ensemble and transfer learning	Breasts cancer Classification	The authors propose an automated screening mammography system which surpasses human experts in predicting breast cancer. Curated on two large datasets - USA and UK - the system shows a reduction of 5.7% and 1.2% in false positives,

					and 9.4% and 2.7% in false negatives, respectively.

US: Ultrasound; **MR**: Magnetic Resonance; **MRI**: Magnetic Resonance Imaging; **DCE-MRI**: Dynamic Contrast Enhancement MRI; **CT**: Computed Tomography; **GBM**: Glioblastoma Multiforme, **CNN**: Convolutional Neural Network**, LN**: Lymph Node; **ILD**: Interstitial Lung Disease; **CMBs**: Cerebral Microbleeds, **FCN**: Fully-Convolutional Network; **FCRN**: Fully-Convolutional Residual Network; **AUC**: Area Under the ROC Curve; **ROC**: Receiver Operating Characteristic

2.3 Data Storage, Management and Sharing in Medical Imaging

Many challenges regarding storing, indexing and data interoperability in medical imaging arise when the data keeps growing. The mere production of large amount of data does not automatically permit the real exploitation of their intrinsic value. The poor curation and semantic annotation of data hinders the training process of machine learning algorithms and integrative analytics. Researchers involved in medical imaging do of course face many more challenges, such as storing, indexing, authorization and privacy issues, data sharing etc. To address these issues a collaboration between research institutions and clinical sites is essential, both at the local and national level. However, with our current knowledge several issues may already be addressed.

First and foremost, the construction of big databases, such as cloud-based databases, is needed to support the massive production of data, as well as to enable the parallel computations of more intricate deep learning methods for data analytics. To ensure reliable databases the principles of the five Big Data "Vs" referring to variability, veracity, volume, velocity, and value must be followed [99]. The provision of metadata is also needed for the descriptive information of the acquired data, and image retrieval since the emergence of content-based image retrieval (CBIR) [100] systems enable now image retrieval by analysing the content of the image instead of, e.g., tags and keywords. To foster multi-institutional collaboration, guided principles, such as FAIR [101], must be followed, which establish a set of recommendations towards making metadata findable, accessible, interoperable, and reusable. Regarding privacy issues, approaches such as K-anonymity [102], L-diversity [103] and T-closeness [104] have been proposed to accomplish the anonymization of medical imaging data. Furthermore, two initiatives QIBA [105] and IBSI [106], aim to reduce variability between different image devices, to standardise the extraction of image biomarkers from acquired imaging, to adopt standardised image biomarker nomenclature, to accelerate the development and adoption of hardware and software standards, and to provide a standardised processing workflow for each imaging modality. Other disease specific initiatives include the Multi-Ethnic Study of Atherosclerosis (MESA), the UK biobank, the Cancer Imaging Archive (TCIA), the Cancer Genome Atlas (TCGA), and the Alzheimer's Disease Neuroimaging Initiative (ADNI).

In clinical sites it is necessary to establish an efficient clinical data repository (CDR). A CDR consolidates data from a variety of clinical sources, and thus allowing clinicians to have an integrated assessment of a single patient. CDR encapsulates data from electronic health (EHR) and medical records (MDR), radiology and pathology archives, tumor registries, biospecimen repositories etc. To ensure the proper classification of the different entities, the adoption of clinical terminology coding (e.g., by SNOMED Clinical Terms®, or the International Classification of Diseases (ICD)) is useful. To ensure the interoperability of a CDR in case of multi-collaboration, it is necessary to rely on standards, such as those defined by the Integrating the Healthcare Enterprise (IHE), HL7 Fast Healthcare Interoperability Resources (FHIR), and Digital Imaging and Communications in Medicine (DICOM). Furthermore, automated ETL ("extract, transform, and load") interfaces and tools allow enterprises to gather data from multiple sources and consolidate it into a single, centralized location [107]. ETL also makes it possible for different types of data to work together

to analyse multi-nodal data in a systematic manner, guide personalized treatment and refine best practices. Finally, integrating software supporting data management process in clinical trials may also be essential to automate all dimensions of the clinical data management process [108].

2.4 Digital Pathology

Single tissue and TMA images can provide important information about diseased tissue and disease mechanics at the sub-cellular scale. Nowadays microscope glass slides can be converted into digital slides with the aid of computer monitor (known as whole slide imaging (WSI), or virtual microscopy) [109], thus enabling to automate the process of analysing single tissue and TMA images and predicting diseases due to the success of machine learning algorithms. FDA approved in 2017 the use of a commercial digital pathology system in clinical settings [110].

2.4.1 Segmentation/Classification and Understanding

In recent years there has been an increased interest in employing computer-aided decision support systems for the proper segmentation and classification of tissue images. However, image processing algorithms must overcome several challenges to extract useful information from tissue images. WSI contain up to thousands of cells and nuclei, and combined by heterogeneity in structure across tissue specimens, they lead to huge datasets. Generalizing the task of nuclear characterization from different tissue phenotypes poses another great challenge, since producing truth tables and IF-THEN datasets for algorithmic training is labour intensive and requires expert knowledge from pathologists [111]. Furthermore, other issues, such as tissue abnormalities across the same specimen, burden the interpretation process which is traditionally made in a subjective manner by multiple pathologists to reach consensus.

Novel methods aim to develop content-based retrieval (CBR) systems for the classification of pathology specimens. These systems enable the automatic search through reference libraries of pathology images based on similar characteristics to a given query [112]. However, the large and high-dimensional datasets can render feature search inefficient. In view of that, hashing techniques, which directly search data without using index structure, can be used to make the data retrieving an efficient process [113]. Deep learning methods have also made it possible to automate many aspects of tissue image processing [114], presenting different methods depending on the malady and the disease site [115]. For example, DL classification methods decide on whether regions of tissue contain tumors, necrosis or immune cells, or if tissue regions area consent to specialist descriptions of tissue patterns. These and other elements like maps of the size, shape, and texture of nuclei in addition with different statistical features may form critical cancer biomarkers [116]. Recently, the most use of various tissue images is targeted on the interaction of cancer with the immune system. E.g., observations that tumor-infiltrating lymphocytes (TILs) are correlated with favourable clinical outcomes like longer disease-free survival in multiple cancer types [117], have led to the development of prognostic scores, such as the Immunoscore [118] and the TIL abundance (TILAB) score [119]. Moreover, a recent study [120] reports on a series of immunogenomic characterizations that include assessments such as total lymphocytic infiltrate, immune cell type fractions, immune gene expression signatures, human leukocyte antigen (HLA) type and expression, neoantigen prediction, T cell and B cell repertoire, and viral RNA expression. The objective evaluation of these biomarkers often poses new challenges. E.g., HLA category A tumor epithelium expression is tough to quantify by eye because of its concurrent presence on tumor and healthy tissue cells [121]. Efforts that tried to combine omics data with pathology images include various statistical and machine learning methods like consensus clustering [122], linear classifier [123], LASSO regression modelling [124], and deep learning [125].

Current tissue diagnostic studies use stained methods, the most common being hematoxylin and eosin (H&E), to provide detailed features of tissue. With the current imaging methods, digital pathology combined with deep learning can formulate relationships unseen by human inspection. E.g., some efforts include the characterization of T cell repertoire in lung cancer [126], the correlation of TIL patterns with molecular data and the generation of tumor infiltrating lymphocyte maps [127]. A common challenge that arises in digital pathology is the batch effect, where tissue slides from different institutions show heterogeneous appearances because of differences in tissue preparation and staining protocols. However, this problem was tackled by the advent of unsupervised learning which allowed to transfer the discriminative information obtained from the in domain to the target domain without requiring relabelling images at the target domain [128]. Finally, DL methods have also been used to detect cancer metastases based on tissue images [129].

2.4.2 Data Management, Querying and Visualization

A common open-source system for data management, feature querying and visualization of whole slide images, is QuIP [130], which utilizes tools like caMicroscope viewer [131] to support the interactive visualization of images and their annotations, and FeatureScape - a visual analytic tool that supports interactive exploration of feature and segmentation maps. Other open-source systems that carry out these or related tasks are QuPath [132], the Pathology Image Informatics Platform (PIIP) [133], the Digital Slide Archive (DSA) [134] and Cytomine [135]. Regarding the format of images, most of recent efforts have libraries like OpenSlide [136] or Bio-Formats [137] to navigate the different formats of plethora of work towards the adoption of a common format (probably that is DICOM format). To handle the ever-increasing datasets and the massive computations to train machine learning algorithms and make predictions, there has been an increased awareness in the use of cloud computing [138], a cost-effective solution for large-scale computing. Software like QuIP includes cloud-based pipelines for tumor infiltrating lymphocyte analysis and nuclear segmentation.

2.5 In Silico Models

2.5.1 Medical Image Reconstruction and Visualization

Medical image reconstruction refers to the 3D surface generation and visualization of different biological components, such as arteries, vessels, organs etc. It also involves mesh generation techniques, followed by rendering techniques used for completing the seamless boundary surface, smoothing and refinement. 3D image reconstruction and visualization enable applications in virtual surgery, neuro-interventions, coronary and aortic stenting etc.

3D medical image reconstruction comprises the following stages: 1) Segmentation of 2D image slices and feature extraction, 2) reconstruction of 3D images, and 3) display of reconstructed images by a corresponding software. The segmentation process has been described in a previous section. Regarding the image reconstruction, note first that different image acquisition methods offer different image information. E.g., CT images offer clear information on bone tissue, while MRI images offer clear information on soft tissue. Thus, to improve the effect of reconstruction, 2D images are first fused. Image fusion refers

to the transformation of different medical images and their spatial coordinates matching. By fusion complementary information is obtained, which can improve the accuracy of clinical diagnosis and treatment. According to [139], the methods of image fusion can be divided into three broad categories:

1. *Spatial Domain*: In this domain, two or more images are directly computed and added in spatial coordinates. This method involves logical computations, such as weighted average, pattern computation and image algebraic calculations.
2. *Changing Domain*: In this domain, an image is first modified and then fused. It includes algorithms such as the Laplacian Pyramid [140], wavelet change [141] etc.
3. *Intelligent Domain*: Concerns the algorithms that try to simulate human being intelligent processing method (i.e., Machine Learning and Deep Learning algorithms).

The selection of applicable meshing and rendering techniques depends on the imaging modality and the corresponding biological element. The earlier reconstruction models were based on handcraft mathematical models. Surface rendering techniques can reconstruct 3D boundaries, like the geometry of arteries and vessels through the iso-contours extracted from 2D image slices [142]. At present, non-uniform rational basis splines (NURBS) is an efficient mathematical model using B-splines to represent curves and surfaces such as aortic, carotid, cerebral and coronary arteries [143]. For the representation of bulk biological elements (e.g., tumors) volume rendering techniques are employed, such as ray-casting [144], light projection field displays [145] and frequency domain transformations (e.g., Fourier, wavelet frames [146]). Other successful handcraft models include the total variation model [147], the Perona-Malik diffusion [148], shock-filters [149], nonlocal methods [150], block matching into 3D data arrays (BM3D) [151] and weighted nuclear norm minimization (WNNM) [152]. These models mostly have solid theoretical foundations and high interpretability. They perform reasonably well in practice, and some of them present the state of the art for certain tasks.

Since 1999, models consisting of data-driven learning strategies and handcraft modelling began to emerge [153]. Compared to purely handcrafted models, these models can better exploit the available data and outperform their corresponding no data-driven counterparts. Meanwhile, the handcrafted framework of the models grants certain interpretability and theoretical foundation to the models. Successful examples include the method of optimal directions [154], the K-SVD [155], learning-based PDE design [156], data-driven tight frame [157], Adaframe [158], low-rank models [159], piecewise-smooth image models [160], and statistical models [161]. In 2012, with the advent of deep learning, various types of CNNs such as ResNet [162] and generative adversarial networks (GANs) [163] were introduced and applied in image re-constructions presenting the current state of the art.

To reap the benefits of reconstruction process and make it available in the analysis of many patient specific cases, automating the process from segmentation to reconstruction is needed. However, this automation process is hindered due to different imaging modalities, varying element geometries, the quality of source images etc. Efficient workflows are also required to segment and reconstruct images. One way to overcome this challenge is the “raw-to-task” workflow as discussed in [153]. Raw-to-task deviates from the traditional workflow where the image analysis is separated into two stages: 1) reconstruction of a high-quality image from raw data, and 2) make a diagnosis based on the high-quality reconstructed image. Raw-to-task unifies the two afore-mentioned stages by connecting two CNNs together and conducting end-to-end training as demonstrated in Fig. 5. Such idea was first introduced in medical imaging by [164].

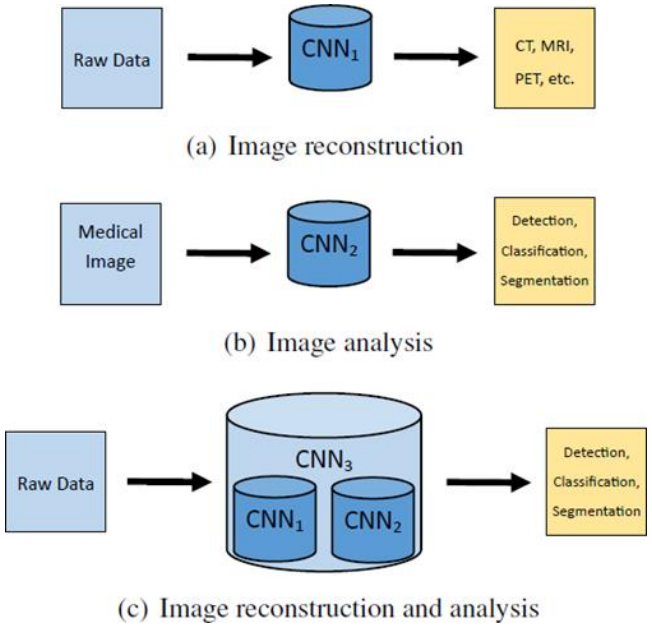

Figure 5. CNN based workflows for medical image reconstruction and analysis [153].

2.5.2 In Silico Modelling of Malignant Tumors

In silico modelling can infer reliable predictions on tumor dynamics, based on several mathematical foundations. Such computational models have been used to investigate the mechanisms that govern cancer progression and invasion, aiming to predict its future spatial and temporal status [165]. Recent efforts move towards multiscale approaches that link the interaction mechanisms at different biological scales [166], e.g., organs' responses, blood dynamics, nutrient transport, and consumption etc., while they are still computationally efficient. Other efforts aim for the development of multi-compartment models which describe the behaviour (i.e., proliferation, migration, etc.) of different cell population [167, 168]. Recent models also try to incorporate the diffusion dynamics and concentration gradients of chemical substances, e.g. oxygen, glucose, drugs etc., and the influence of each cell expression resulting from intracellular signalling pathways and gene expressions. However so far, the capabilities of these models are limited by the inability to simulate cellular interactions (e.g. cell to cell adhesion) and sub-cellular chain processes [95], which determine cellular behaviour.

2.5.3 Digital Twins

A digital twin (DT) is a digital replica of a physical asset [169], trying to mimic its behaviour based on mathematical models. Thus, by real-time monitoring of the physical asset, DTs can evaluate different parameters of interest of its current state and even make predictions for the future. Although this concept initially emerged in industry and aerospace sector, its capabilities can also extend to healthcare.

DT comprises a physical asset, its digital replica, and a bidirectional synchronized relation between the two. Other desirable components include 1) IoT devices[2] for the collection of data from the physical asset, 2) an integrated workflow to gather the data from different IoT devices, monitor the physical system and

[2] IoT devices refer to anything that has a sensor attached to it and can transmit data with the help of the internet.

provide input to machine learning methods to make predictions and give feedback to guarantee the correct behavior for the physical system, 3) big data analysis and storage tools, 4) protocols to ensure the security and fidelity of the data, and 5) the standardization of evaluation tests to ensure, e.g., the accuracy and robustness of the model.

DT may be a useful tool for personalized medicine. FDA deemed medication ineffective for 38-75% of patients with common maladies [170], resulting in patient suffering and an increase in healthcare costs. Conventional medication relies on several biomarkers and its limited effectiveness often emanates from the altered interactions of genes that differ among patients with the same diagnosis. DTs can help to increase specificity of an administrated drug and refine methods to test its effectiveness more quickly and economically [171]. Swedish Digital Twin Consortium (SDTC) [172] is an initiative moving towards this direction. SDTC aims to 1) construct many digital twins of network models of all molecular, phenotypic, and other factors relative to disease dynamics in each individual, and 2) to computationally test with different drugs those digital twins to identify the best performing drug for each patient (Fig. 6). Recent work [173] presents methods and algorithms aiming at synthesizing optimal personalized treatments by means of In Silico Clinical Trials (ISCT), by exploiting quantitative models of the physiology and drugs Pharmacokinetics/Pharmacodynamics (PKPD) of interest, and clinical measurements on human patients from which they defined their digital twins. Furthermore, digital twins of organs, such as the liver [174], the heart [175] and kidney [176], have been constructed which combine various functional measurements with multi-scale modelling. Finally, the DT concept is also proposed in clinical healthcare for trauma management (i.e., procedures, administered drugs, diagnostics reports, vital signs etc.) [169], as well as for cancer preclinical investigation [177].

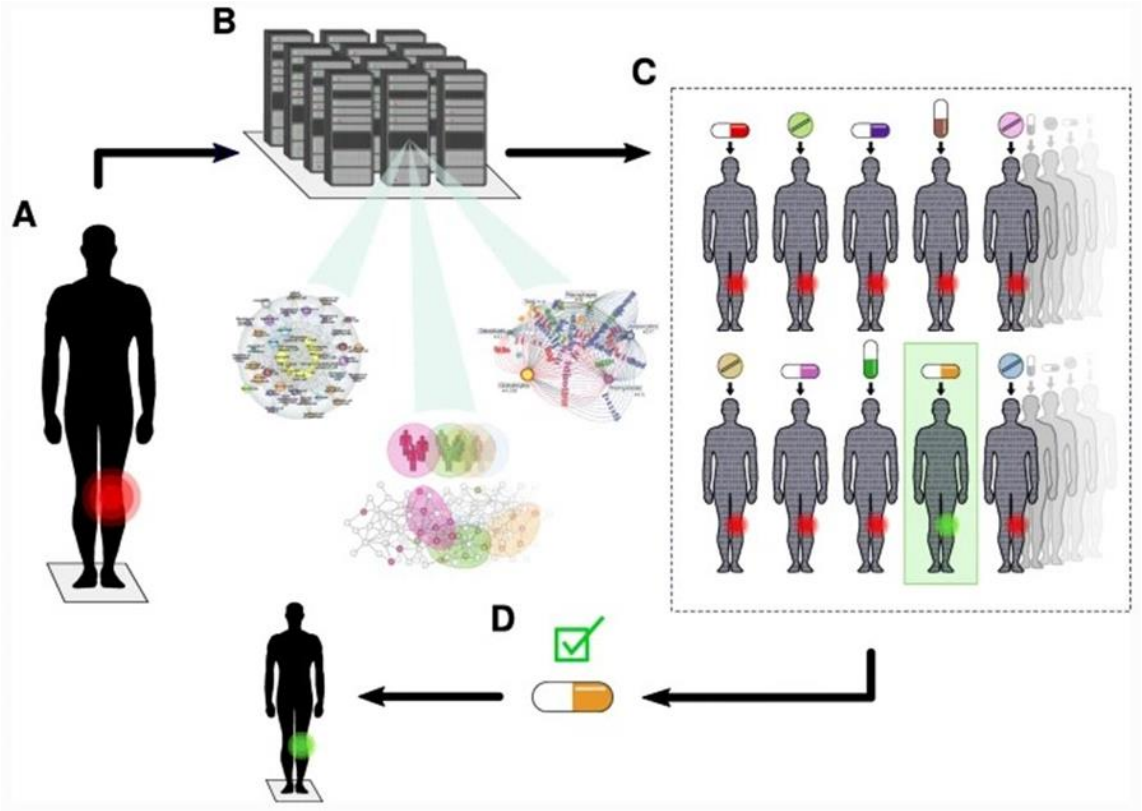

Figure 6. The digital twin concept for personalized medicine: **A.** An individual patient has a local sign of disease (red). **B**. A digital twin of this patient is constructed in unlimited copies, based on computational network models

of thousands of disease-relevant variables. **C.** Each twin is computationally treated with one or more of the thousands of drugs. This results in digital cure of one patient (green). **D.** The drug that has the best effect on the digital twin is selected for treatment of the patient [171].

2.6 Concluding Remarks & Future Directions

In this section we first introduced the image acquisition techniques from the most common, routine techniques (e.g., MRI, CT) to the new forthcoming techniques (e.g., Ultrasound enriched with bubbles and Super-resolution microscopy techniques) applied in medical imaging. Next, we described the phase of image post-processing following the image acquisition. We said that this phase involves the segmentation and classification of images. We mentioned the most common, and yet powerful techniques in image segmentation, including the machine and deep learning approaches. Next, we presented the stages of classification process, as well as the techniques applied in each stage. In fact, image post-processing is more complicated than the way it is covered here, though we highlighted the basic and most prevalent concepts, as well as the current state-of-the-art algorithms and trends. In the following, we mentioned the challenges that arise due to the large volume of data produced and the proposed solutions of imaging informatics community. We discussed the contribution of imaging informatics to the advent of digital pathology, describing the post-processing of tissue images, the interpretation of the results for proper diagnostics, and the platforms that allow for data management, visualization and processing in this field. Then, we presented the most profound applications of imaging informatics in healthcare, including the digital 3D reconstruction and visualization of different anatomical sites, and new forthcoming concepts such as the in-silico modeling of malignant tumors and the concept of digital twins.

As regards the future directions, the adoption of GPU (graphics processing unit) in medical physics seems to emerge [178, 179]. Although originally designed for accelerating the production of computer graphics, the GPU has emerged as a versatile platform for running massively parallel computations. The introduction of new multi-core architectures of the GPU and the advent of programmable GPUs by the non-expert, along with computer-oriented GPU interfaces lead to advancements in medical imaging that allow, e.g., the real time reconstruction of an image or the ability to handle large data sets from multiple IoT devices, underpinning the concept of Digital Twins.

Regarding data management, the ever-increasing datasets of imaging data, complemented by EMR and EHR, -omics, and other physiological data will pose challenges like data fidelity and integration from various imaging sources, querying, data analysis, storage, interoperability, security and privacy issues. Currently, deep learning methods have a dominant role in image processing for presenting high accuracy in classification/segmentation tasks and a reasonably good performance at producing synthetic images from different acquisition techniques. The advent of GANs, since they were first devised in 2014 by Goodfellow et al. [163], allow now the models to train on unlabeled data, learn messy and complicated distributions of the data, and generate data that is similar to real data. In the future efforts like fine-tuning and transfer learning aim also to create methods that rely on smaller datasets and still generalize well. There is an ongoing interest in advancing DL methods across a wide spectrum of healthcare data, from EHR [180], genomics [181] and other physiological parameters [182]. Other emerging trends include natural language data processing [183], where computers analyze large amounts of natural language data and thus enable applications such as verbal querying.

Emerging radiogenomics investigate the association between various imaging phenotypes and -omics data, e.g. a tumor's texture and size characteristics in terms of molecular profiling, and thus providing new insights for disease aetiology, dynamics, and response to a treatment.

Finally, to accelerate new knowledge discovery, there must be more initiatives that include multi-institutional collaboration and broader networks of research groups, standardization of workflows, open-access datasets from well-annotated large cohorts, reproducible and well explicable research studies.

2.7 References

[1] J. A. Jensen, *Estimation of blood velocities using ultrasound: a signal processing approach*. Cambridge University Press, 1996.

[2] P. N. Burns, P. Hilpert, and B. B. Goldberg, "Intravenous contrast agent for ultrasound Doppler: In vivo measurement of small tumor vessel dose-response," in *[1990] Proceedings of the Twelfth Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1990, pp. 322–324.

[3] J. R. Lindner, "Microbubbles in medical imaging: current applications and future directions," *Nat. Rev. Drug Discov.*, vol. 3, no. 6, pp. 527–533, 2004.

[4] J.-L. Gennisson, T. Deffieux, M. Fink, and M. Tanter, "Ultrasound elastography: principles and techniques," *Diagn. Interv. Imaging*, vol. 94, no. 5, pp. 487–495, 2013.

[5] N. J. Bravo-Valenzuela *et al.*, "Fetal cardiac function by three-dimensional ultrasound using 4D-STIC and VOCAL–an update," *J. Ultrason.*, vol. 19, no. 79, p. 287, 2019.

[6] K. Christensen-Jeffries *et al.*, "Super-resolution ultrasound imaging," *Ultrasound Med. Biol.*, vol. 46, no. 4, pp. 865–891, 2020.

[7] J. Keyriläinen, A. Bravin, M. Fernández, M. Tenhunen, P. Virkkunen, and P. Suortti, "Phase-contrast X-ray imaging of breast," *Acta radiol.*, vol. 51, no. 8, pp. 866–884, 2010.

[8] C. H. McCollough, S. Leng, L. Yu, and J. G. Fletcher, "Dual-and multi-energy CT: principles, technical approaches, and clinical applications," *Radiology*, vol. 276, no. 3, pp. 637–653, 2015.

[9] J. White, G. Couzens, and C. Jeffery, "The use of 4D-CT in assessing wrist kinematics and pathology: a narrative review," *Bone Joint J.*, vol. 101, no. 11, pp. 1325–1330, 2019.

[10] S. D. Rawson, J. Maksimcuka, P. J. Withers, and S. H. Cartmell, "X-ray computed tomography in life sciences," *BMC Biol.*, vol. 18, no. 1, pp. 1–15, 2020.

[11] M. P. Hartung, T. M. Grist, and C. J. François, "Magnetic resonance angiography: current status and future directions," *J. Cardiovasc. Magn. Reson.*, vol. 13, no. 1, pp. 1–11, 2011.

[12] K.-D. Merboldt, W. Hanicke, and J. Frahm, "Self-diffusion NMR imaging using stimulated echoes," *J. Magn. Reson.*, vol. 64, no. 3, pp. 479–486, 1985.

[13] L. J. O'Donnell and C.-F. Westin, "An introduction to diffusion tensor image analysis," *Neurosurg. Clin.*, vol. 22, no. 2, pp. 185–196, 2011.

[14] F. J. Al Badarin and S. Malhotra, "Diagnosis and prognosis of coronary artery disease with SPECT and PET," *Curr. Cardiol. Rep.*, vol. 21, no. 7, pp. 1–11, 2019.

[15] I. R. Yoo, "Bone SPECT/CT of the Foot and Ankle: Potential Clinical Application for Chronic Foot Pain," *Nucl. Med. Mol. Imaging (2010).*, vol. 54, no. 1, pp. 1–8, 2020.

[16] E. Bergeron, E. Désilets, X. V. Do, D. McNamara, S. Chergui, and M. Bensoussan, "A case of torsion of the gallbladder suspected with SPECT-CT: review and recommendations," *Case Rep. Surg.*, vol. 2020, 2020.

[17] E. Grady, "Gastrointestinal bleeding scintigraphy in the early 21st century," *J. Nucl. Med.*, vol. 57, no. 2, pp. 252–259, 2016.

[18] L. Zhu, K. Ploessl, and H. F. Kung, "PET/SPECT imaging agents for neurodegenerative diseases," *Chem. Soc. Rev.*, vol. 43, no. 19, pp. 6683–6691, 2014.

[19] T. Pan *et al.*, "Elevated expression of glutaminase confers glucose utilization via glutaminolysis in prostate cancer," *Biochem. Biophys. Res. Commun.*, vol. 456, no. 1, pp. 452–458, 2015.

[20] H. Feng *et al.*, "Nuclear imaging of glucose metabolism: beyond 18F-FDG," *Contrast Media Mol. Imaging*, vol. 2019, 2019.

[21] P. Duke, *Synchrotron radiation: production and properties*, vol. 3. Oxford University Press, 2009.

[22] P. Suortti and W. Thomlinson, "Medical applications of synchrotron radiation," *Phys. Med. Biol.*, vol. 48, no. 13, p. R1, 2003.

[23] J. A. LAISSUE, J. F. LE BASZ, G. LE DUCI, N. LYUBIMOVA, C. NEMOZ, and M. RENIERI, "RESEARCH AT THE EUROPEAN SYN CHROTRON RADIATION FACILITY NIEDICAL BEAMLINE," *Cell. Mol. Biol.*, vol. 46, no. 6, pp. 1053–1063, 2000.

[24] "A short introduction to light electron microscopy." 2015, [Online]. Available: http://www.zmb.uzh.ch/static/toolbox/assets/Script_2015_Toolbox.pdf.

[25] I. Rakotoson *et al.*, "Fast 3-D imaging of brain organoids with a new single-objective planar-illumination two-photon microscope," *Front. Neuroanat.*, vol. 13, p. 77, 2019.

[26] S. Luro, L. Potvin-Trottier, B. Okumus, and J. Paulsson, "Isolating live cells after high-throughput, long-term, time-lapse microscopy," *Nat. Methods*, vol. 17, no. 1, pp. 93–100, 2020.

[27] C. Ricard *et al.*, "Two-photon probes for in vivo multicolor microscopy of the structure and signals of brain cells," *Brain Struct. Funct.*, vol. 223, no. 7, pp. 3011–3043, 2018.

[28] S. W. Perry, R. M. Burke, and E. B. Brown, "Two-photon and second harmonic microscopy in clinical and translational cancer research," *Ann. Biomed. Eng.*, vol. 40, no. 2, pp. 277–291, 2012.

[29] F. Helmchen and W. Denk, "Deep tissue two-photon microscopy," *Nat. Methods*, vol. 2, no. 12, pp. 932–940, 2005.

[30] J. E. Sleeman, "Dynamics of the mammalian nucleus: can microscopic movements help us to understand our genes?," *Philos. Trans. R. Soc. London. Ser. A Math. Phys. Eng. Sci.*, vol. 362, no. 1825, pp. 2775–2793, 2004.

[31] C. Lal and M. J. Leahy, "An updated review of methods and advancements in microvascular blood flow imaging," *Microcirculation*, vol. 23, no. 5, pp. 345–363, 2016.

[32] C. A. Franco *et al.*, "Dynamic endothelial cell rearrangements drive developmental vessel regression," *PLoS Biol*, vol. 13, no. 4, p. e1002125, 2015.

[33] C. Saldanha, "Instrumental analysis applied to erythrocyte properties," *J. Cell. Biotechnol.*, vol. 1, no. 1, pp. 81–93, 2015.

[34] G. Pellacani, P. Pepe, A. Casari, and C. Longo, "Reflectance confocal microscopy as a second-level examination in skin oncology improves diagnostic accuracy and saves unnecessary excisions: a longitudinal prospective study," *Br. J. Dermatol.*, vol. 171, no. 5, pp. 1044–1051, 2014.

[35] E. Isasi, L. Barbeito, and S. Olivera-Bravo, "Increased blood–brain barrier permeability and alterations in perivascular astrocytes and pericytes induced by intracisternal glutaric acid," *Fluids*

Barriers CNS, vol. 11, no. 1, pp. 1–12, 2014.

[36] E. Cinotti *et al.*, "Quantification of capillary blood cell flow using reflectance confocal microscopy," *Ski. Res. Technol.*, vol. 20, no. 3, pp. 373–378, 2014.

[37] M. Venturini *et al.*, "In vivo reflectance confocal microscopy features of cutaneous microcirculation and epidermal and dermal changes in diffuse systemic sclerosis and correlation with histological and videocapillaroscopic findings," *Eur. J. Dermatology*, vol. 24, no. 3, pp. 349–355, 2014.

[38] J. M. Guerra, "Photon tunneling microscopy," *Appl. Opt.*, vol. 29, no. 26, pp. 3741–3752, 1990.

[39] Wikipedia, "Near-field scanning optical microscope." https://en.wikipedia.org/wiki/Near-field_scanning_optical_microscope.

[40] S. van Deventer, A. B. Arp, and A. B. van Spriel, "Dynamic Plasma Membrane Organization: A Complex Symphony," *Trends Cell Biol.*, 2020.

[41] B. Traenkle, S. Segan, F. O. Fagbadebo, P. D. Kaiser, and U. Rothbauer, "A novel epitope tagging system to visualize and monitor antigens in live cells with chromobodies," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020.

[42] J. K. Eykelenboom and T. U. Tanaka, "Zooming in on chromosome dynamics," *Cell Cycle*, vol. 19, no. 12, pp. 1422–1432, 2020.

[43] H. Steffens *et al.*, "Chronic in vivo STED nanoscopy uncovers multiple drivers of shape volatility in stable cortical spines," *bioRxiv*, 2020.

[44] M. Lakadamyali and M. P. Cosma, "Visualizing the genome in high resolution challenges our textbook understanding," *Nat. Methods*, vol. 17, no. 4, pp. 371–379, 2020.

[45] B. Apter *et al.*, "Fluorescence Phenomena in Amyloid and Amyloidogenic Bionanostructures," *Crystals*, vol. 10, no. 8, p. 668, 2020.

[46] J. C. Boatz, T. Piretra, A. Lasorsa, I. Matlahov, J. F. Conway, and P. C. A. van der Wel, "Protofilament structure and supramolecular polymorphism of aggregated mutant huntingtin exon 1," *J. Mol. Biol.*, vol. 432, no. 16, pp. 4722–4744, 2020.

[47] R. Erni, M. D. Rossell, C. Kisielowski, and U. Dahmen, "Atomic-resolution imaging with a sub-50-pm electron probe," *Phys. Rev. Lett.*, vol. 102, no. 9, p. 96101, 2009.

[48] M. Matyszewski and J. Sohn, "Preparation of filamentous proteins for electron microscopy visualization and reconstruction," *Methods Enzymol.*, vol. 625, pp. 167–176, 2019.

[49] J. F. Bille, "Optical Coherence Tomography (OCT): Principle and Technical Realization--High Resolution Imaging in Microscopy and Ophthalmology: New Frontiers in Biomedical Optics," 2019.

[50] R. F. Spaide, J. G. Fujimoto, N. K. Waheed, S. R. Sadda, and G. Staurenghi, "Optical coherence tomography angiography," *Prog. Retin. Eye Res.*, vol. 64, pp. 1–55, 2018.

[51] I. Abd El-Sadek *et al.*, "Optical coherence tomography-based tissue dynamics imaging for longitudinal and drug response evaluation of tumor spheroids," *Biomed. Opt. Express*, vol. 11, no. 11, pp. 6231–6248, 2020.

[52] A. Ramier *et al.*, "In vivo measurement of shear modulus of the human cornea using optical coherence elastography," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020.

[53] A. B. E. Attia *et al.*, "A review of clinical photoacoustic imaging: Current and future trends,"

Photoacoustics, vol. 16, p. 100144, 2019.

[54] M. Heijblom *et al.*, "Photoacoustic image patterns of breast carcinoma and comparisons with Magnetic Resonance Imaging and vascular stained histopathology," *Sci. Rep.*, vol. 5, no. 1, pp. 1–16, 2015.

[55] M. Yang *et al.*, "Photoacoustic/ultrasound dual imaging of human thyroid cancers: an initial clinical study," *Biomed. Opt. Express*, vol. 8, no. 7, pp. 3449–3457, 2017.

[56] J. Jo *et al.*, "A functional study of human inflammatory arthritis using photoacoustic imaging," *Sci. Rep.*, vol. 7, no. 1, pp. 1–9, 2017.

[57] Y. Zhu *et al.*, "Light emitting diodes based photoacoustic imaging and potential clinical applications," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018.

[58] "AcousticX, Cyberdyne Inc., Tokyo, Japan." https://www.cyberdyne.jp/english/products/pa01.html.

[59] Tomáš Čižmár, "Exploiting multimode waveguides for in vivo imaging." 2015, [Online]. Available: https://spie.org/news/6106-exploiting-multimode-waveguides-for-in-vivo-imaging?SSO=1.

[60] A. S. Dar and D. Padha, "Medical image segmentation: a review of recent techniques, advancements and a comprehensive comparison"," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 7, pp. 114–124, 2019.

[61] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imaging*, vol. 13, no. 1, pp. 146–165, 2004.

[62] S. Gould, T. Gao, and D. Koller, "Region-based Segmentation and Object Detection.," in *NIPS*, 2009, vol. 1, p. 2.

[63] D. Kelkar and S. Gupta, "Improved quadtree method for split merge image segmentation," in *2008 First International Conference on Emerging Trends in Engineering and Technology*, 2008, pp. 44–47.

[64] A. Anand, S. S. Tripathy, and R. S. Kumar, "An improved edge detection using morphological Laplacian of Gaussian operator," in *2015 2nd International conference on signal processing and integrated networks (SPIN)*, 2015, pp. 532–536.

[65] G. N. Chaple, R. D. Daruwala, and M. S. Gofane, "Comparisions of Robert, Prewitt, Sobel operator based edge detection methods for real time uses on FPGA," in *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, 2015, pp. 1–4.

[66] M. Sharma, G. N. Purohit, and S. Mukherjee, "Information retrieves from brain MRI images for tumor detection using hybrid technique K-means and artificial neural network (KMANN)," in *Networking communication and data knowledge engineering*, Springer, 2018, pp. 145–157.

[67] A. Hanbury and H. Müller, "VISCERAL: Evaluation-as-a-Service for Medical Imaging," in *Cloud-Based Benchmarking of Medical Image Analysis*, Springer, Cham, 2017, pp. 3–13.

[68] T. Heimann and H. Delingette, "Model-based segmentation," in *Biomedical Image Processing*, Springer, 2010, pp. 279–303.

[69] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.

[70] D. Metaxas and D. Terzopoulos, "Constrained deformable superquadrics and nonrigid motion

tracking," in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1991, pp. 337–338.

[71] N. Padmasini, R. Umamaheswari, and M. Y. Sikkandar, "State-of-the-Art of Level-Set Methods in Segmentation and Registration of Spectral Domain Optical Coherence Tomographic Retinal Images," *Soft Comput. Based Med. Image Anal.*, pp. 163–181, 2018.

[72] T. F. Chan and L. A. Vese, "Active contour and segmentation models using geometric PDE's for medical imaging," in *Geometric methods in bio-medical image processing*, Springer, 2002, pp. 63–75.

[73] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations," *J. Comput. Phys.*, vol. 79, no. 1, pp. 12–49, 1988.

[74] W. Roetzel, X. Luo, and D. Chen, "Optimal design of heat exchanger networks," in *Design and Operation of Heat Exchangers and Their Networks*, Elsevier, 2019, pp. 231–318.

[75] H. Kotadiya and D. Patel, "Review of medical image classification techniques," in *Third International Congress on Information and Communication Technology*, 2019, pp. 361–369.

[76] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[77] J. Liu, J. Li, T. Liu, and J. Tam, "Graded Image Generation Using Stratified CycleGAN," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 760–769.

[78] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Med. Image Anal.*, vol. 52, pp. 128–143, 2019.

[79] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[80] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, 2016, pp. 424–432.

[81] A. A. A. Setio *et al.*, "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[82] J. Ding, A. Li, Z. Hu, and L. Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 559–567.

[83] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2014.

[84] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014, pp. 818–833.

[85] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[86] "MICCAI." http://www.miccai.org/.

[87] "CAMELYON." https://camelyon17.grand-challenge.org/.

[88] "IEEE brain." https://brain.ieee.org/2020-competitions-and-challenges/.

[89] A. S. Panayides *et al.*, "Ai in medical imaging informatics: Current challenges and future directions," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 7, pp. 1837–1857, 2020.

[90] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[91] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1182–1195, 2016.

[92] K. Lekadir *et al.*, "A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound," *IEEE J. Biomed. Heal. informatics*, vol. 21, no. 1, pp. 48–55, 2016.

[93] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imaging*, vol. 36, no. 4, pp. 994–1004, 2016.

[94] J. Lao *et al.*, "A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme," *Sci. Rep.*, vol. 7, no. 1, pp. 1–8, 2017.

[95] L. Oakden-Rayner, G. Carneiro, T. Bessen, J. C. Nascimento, A. P. Bradley, and L. J. Palmer, "Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework," *Sci. Rep.*, vol. 7, no. 1, pp. 1–13, 2017.

[96] Z. Zhu, E. Albadawy, A. Saha, J. Zhang, M. R. Harowicz, and M. A. Mazurowski, "Deep learning for identifying radiogenomic associations in breast cancer," *Comput. Biol. Med.*, vol. 109, pp. 85–90, 2019.

[97] A. Chartsias *et al.*, "Factorised spatial representation learning: Application in semi-supervised myocardial segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 490–498.

[98] S. M. McKinney *et al.*, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[99] C. S. Mayo *et al.*, "The big data effort in radiation oncology: Data mining or data farming?," *Adv. Radiat. Oncol.*, vol. 1, no. 4, pp. 260–271, 2016.

[100] N. F. Haq, M. Moradi, and Z. J. Wang, "A deep community based approach for large scale content based X-ray image retrieval," *Med. Image Anal.*, vol. 68, p. 101847, 2021.

[101] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. data*, vol. 3, no. 1, pp. 1–9, 2016.

[102] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[103] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. from Data*, vol. 1, no. 1, pp. 3-es, 2007.

[104] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, 2007, pp. 106–115.

[105] "Quantitative Imaging Biomarkers Alliance (QIBA)." https://www.rsna.org/research/quantitative-imaging-biomarkers-alliance.

[106] "The image biomarker standardisation initiative (IBSI)." https://ibsi.readthedocs.io/en/latest/.

[107] P. Pellegrini and A. Grasso, "System and method for automating ETL application." Google Patents, Sep. 20, 2011.

[108] A. Nourani, H. Ayatollahi, and M. S. Dodaran, "A review of clinical data management systems used in clinical trials," *Rev. Recent Clin. Trials*, vol. 14, no. 1, pp. 10–23, 2019.

[109] L. Pantanowitz, A. Sharma, A. B. Carter, T. Kurc, A. Sussman, and J. Saltz, "Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives," *J. Pathol. Inform.*, vol. 9, 2018.

[110] A. J. Evans *et al.*, "US Food and Drug Administration approval of whole slide imaging for primary diagnosis: a key milestone is reached and new questions are raised," *Arch. Pathol. Lab. Med.*, vol. 142, no. 11, pp. 1383–1387, 2018.

[111] T. Hayakawa, V. B. S. Prasath, H. Kawanaka, B. J. Aronow, and S. Tsuruoka, "Computational nuclei segmentation methods in digital pathology: a survey," *Arch. Comput. Methods Eng.*, pp. 1–13, 2019.

[112] M. Jiang, S. Zhang, J. Huang, L. Yang, and D. N. Metaxas, "Joint kernel-based supervised hashing for scalable histopathological image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 366–373.

[113] X. Zhang, F. Xing, H. Su, L. Yang, and S. Zhang, "High-throughput histopathological image analysis via robust cell segmentation and hashing," *Med. Image Anal.*, vol. 26, no. 1, pp. 306–315, 2015.

[114] S. Deng *et al.*, "Deep learning in digital pathology image analysis: a survey," *Front. Med.*, pp. 1–18, 2020.

[115] Q. D. Vu *et al.*, "Methods for segmentation and classification of digital microscopy tissue images," *Front. Bioeng. Biotechnol.*, vol. 7, p. 53, 2019.

[116] C. S. Herrington, R. Poulsom, and P. J. Coates, "Recent advances in pathology: the 2020 annual review issue of the journal of pathology," *J. Pathol.*, vol. 250, no. 5, pp. 475–479, 2020.

[117] B. Mlecnik, G. Bindea, F. Pagès, and J. Galon, "Tumor immunosurveillance in human cancers," *Cancer Metastasis Rev.*, vol. 30, no. 1, pp. 5–12, 2011.

[118] D. Bruni, H. K. Angell, and J. Galon, "The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy," *Nat. Rev. Cancer*, vol. 20, no. 11, pp. 662–680, 2020.

[119] M. Shaban *et al.*, "A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, 2019.

[120] V. Thorsson *et al.*, "The immune landscape of cancer," *Immunity*, vol. 48, no. 4, pp. 812–830, 2018.

[121] D. Krijgsman *et al.*, "A method for semi-automated image analysis of HLA class I tumour epithelium expression in rectal cancer," *Eur. J. Histochem. EJH*, vol. 63, no. 2, 2019.

[122] C. Wang, R. Machiraju, and K. Huang, "Breast cancer patient stratification using a molecular regularized consensus clustering method," *Methods*, vol. 67, no. 3, pp. 304–312, 2014.

[123] Y. Yuan *et al.*, "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling," *Sci. Transl. Med.*, vol. 4, no. 157, pp. 157ra143-157ra143, 2012.

[124] C. Wang, H. Su, L. Yang, and K. Huang, "Integrative analysis for lung adenocarcinoma predicts morphological features associated with genetic variations," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, 2017, pp. 82–93.

[125] J. N. Kather *et al.*, "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nat. Med.*, vol. 25, no. 7, pp. 1054–1056, 2019.

[126] A. Reuben *et al.*, "Comprehensive T cell repertoire characterization of non-small cell lung cancer," *Nat. Commun.*, vol. 11, no. 1, pp. 1–13, 2020.

[127] J. Saltz *et al.*, "Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images," *Cell Rep.*, vol. 23, no. 1, pp. 181–193, 2018.

[128] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, "Adversarial domain adaptation for classification of prostate histopathology whole-slide images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 201–209.

[129] C. Pan *et al.*, "Deep learning reveals cancer metastasis and therapeutic antibody targeting in the entire body," *Cell*, vol. 179, no. 7, pp. 1661–1676, 2019.

[130] J. Saltz *et al.*, "A containerized software system for generation, management, and exploration of features from whole slide tissue images," *Cancer Res.*, vol. 77, no. 21, pp. e79–e82, 2017.

[131] "CAMICROSCOPE." https://wolf.cci.emory.edu//camic_org/apps/landing/landing.html.

[132] P. Bankhead *et al.*, "QuPath: Open source software for digital pathology image analysis," *Sci. Rep.*, vol. 7, no. 1, pp. 1–7, 2017.

[133] A. L. Martel *et al.*, "An image analysis resource for cancer research: PIIP—pathology image informatics platform for visualization, analysis, and management," *Cancer Res.*, vol. 77, no. 21, pp. e83–e86, 2017.

[134] D. A. Gutman *et al.*, "The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research," *Cancer Res.*, vol. 77, no. 21, pp. e75–e78, 2017.

[135] R. Marée *et al.*, "Cytomine: An open-source software for collaborative analysis of whole-slide images," *Diagn. Pathol.*, vol. 1, no. 8, 2016.

[136] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "OpenSlide: A vendor-neutral software foundation for digital pathology," *J. Pathol. Inform.*, vol. 4, 2013.

[137] J. Moore *et al.*, "OMERO and Bio-Formats 5: flexible access to large bioimaging datasets at scale," in *Medical Imaging 2015: Image Processing*, 2015, vol. 9413, p. 941307.

[138] E. Abels *et al.*, "Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association," *J. Pathol.*, vol. 249, no. 3, pp. 286–294, 2019.

[139] L. Zheng, G. Li, and J. Sha, "The survey of medical image 3D reconstruction," in *Fifth International Conference on Photonics and Imaging in Biology and Medicine*, 2007, vol. 6534, p. 65342K.

[140] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," in *Readings in computer vision*, Elsevier, 1987, pp. 671–679.

[141] L. I. U. G. X. Y. W. Hai, "A wavelet-decomposition-based image fusion scheme and its performance evaluation," 自动化学报, vol. 6, 2002.

[142] L. Athanasiou *et al.*, "Three-dimensional reconstruction of coronary arteries and plaque morphology using CT angiography–comparison and registration with IVUS," *BMC Med. Imaging*, vol. 16, no. 1, pp. 1–13, 2016.

[143] F. Galassi *et al.*, "3D reconstruction of coronary arteries from 2D angiographic projections using non-uniform rational basis splines (NURBS) for accurate modelling of coronary stenoses," *PLoS One*, vol. 13, no. 1, p. e0190650, 2018.

[144] L. J. Deakin and M. A. Knackstedt, "Efficient ray casting of volumetric images using distance maps for empty space skipping," *Comput. Vis. Media*, vol. 6, no. 1, pp. 53–63, 2020.

[145] P. A. Kara *et al.*, "Perceptual quality of reconstructed medical images on projection-based light field displays," in *eHealth 360°*, Springer, 2017, pp. 476–483.

[146] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999.

[147] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D nonlinear Phenom.*, vol. 60, no. 1–4, pp. 259–268, 1992.

[148] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, 1990.

[149] S. Osher and L. I. Rudin, "Feature-oriented image enhancement using shock filters," *SIAM J. Numer. Anal.*, vol. 27, no. 4, pp. 919–940, 1990.

[150] A. Buades, B. Coll, and J.-M. Morel, "Image denoising methods. A new nonlocal principle," *SIAM Rev.*, vol. 52, no. 1, pp. 113–147, 2010.

[151] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.

[152] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2862–2869.

[153] H.-M. Zhang and B. Dong, "A review on deep learning in medical image reconstruction," *J. Oper. Res. Soc. China*, pp. 1–30, 2020.

[154] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, 1999, vol. 5, pp. 2443–2446.

[155] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[156] R. Liu, Z. Lin, W. Zhang, and Z. Su, "Learning PDEs for image restoration via optimal control," in *European Conference on Computer Vision*, 2010, pp. 115–128.

[157] C. Bao, H. Ji, and Z. Shen, "Convergence analysis for iterative data-driven tight frame construction scheme," *Appl. Comput. Harmon. Anal.*, vol. 38, no. 3, pp. 510–523, 2015.

[158] C. Tai and E. Weinan, "Multiscale adaptive representation of signals: I. the basic framework," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 4875–4912, 2016.

[159] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization," in *Proceedings of the IEEE*

conference on computer vision and pattern recognition, 2016, pp. 5249–5257.

[160] D. B. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Commun. pure Appl. Math.*, 1989.

[161] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: a review," *Med. Image Anal.*, vol. 13, no. 4, pp. 543–563, 2009.

[162] X. Yu, Z. Yu, and S. Ramalingam, "Learning strict identity mappings in deep residual networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4432–4440.

[163] I. J. Goodfellow *et al.*, "Generative adversarial networks," *arXiv Prepr. arXiv1406.2661*, 2014.

[164] D. Wu, K. Kim, B. Dong, G. El Fakhri, and Q. Li, "End-to-end lung nodule detection in computed tomography," in *International workshop on machine learning in medical imaging*, 2018, pp. 37–45.

[165] L. C. Franssen, T. Lorenzi, A. E. F. Burgess, and M. A. J. Chaplain, "A mathematical framework for modelling the metastatic spread of cancer," *Bull. Math. Biol.*, vol. 81, no. 6, pp. 1965–2010, 2019.

[166] Z. Wang and P. K. Maini, "Editorial special section on multiscale cancer modeling," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 3, pp. 501–503, 2017.

[167] M. Papadogiorgaki, P. Koliou, and M. E. Zervakis, "Glioma growth modeling based on the effect of vital nutrients and metabolic products," *Med. Biol. Eng. Comput.*, vol. 56, no. 9, pp. 1683–1697, 2018.

[168] G. B. Machiraju, P. Mallick, and H. B. Frieboes, "Multicompartment modeling of protein shedding kinetics during vascularized tumor growth," *Sci. Rep.*, vol. 10, no. 1, pp. 1–16, 2020.

[169] A. Croatti, M. Gabellini, S. Montagna, and A. Ricci, "On the Integration of Agents and Digital Twins in Healthcare," *J. Med. Syst.*, vol. 44, no. 9, pp. 1–8, 2020.

[170] U. S. FDA, "Paving the way for personalized medicine," *FDA's Role a new Era Med. Prod. Dev. US Dep. Heal. Hum. Serv.*, pp. 1–61, 2013.

[171] B. Björnsson *et al.*, "Digital twins to personalize medicine," *Genome Med.*, vol. 12, no. 1, pp. 1–4, 2020.

[172] "Swedish Digital Twin Consortium." https://www.sdtc.se/.

[173] S. Sinisi, V. Alimguzhin, T. Mancini, E. Tronci, F. Mari, and B. Leeners, "Optimal personalised treatment computation through in silico clinical trials on patient digital twins," *Fundam. Informaticae*, vol. 174, no. 3–4, pp. 283–310, 2020.

[174] K. Subramanian, "Digital Twin for Drug Discovery and Development—The Virtual Liver," *J. Indian Inst. Sci.*, pp. 1–10, 2020.

[175] T. D. Nguyen, O. E. Kadri, and R. S. Voronov, "An Introductory Overview of Image-Based Computational Modeling in Personalized Cardiovascular Medicine," *Front. Bioeng. Biotechnol.*, vol. 8, 2020.

[176] P. J. Harris *et al.*, "The Virtual Kidney: an eScience interface and Grid portal," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 367, no. 1896, pp. 2141–2159, 2009.

[177] M. Di Filippo *et al.*, "Single-cell digital twins for cancer preclinical investigation," in *Metabolic Flux Analysis in Eukaryotic Cells*, Springer, 2020, pp. 331–343.

[178] P. Després and X. Jia, “A review of GPU-based medical image reconstruction,” *Phys. Medica*, vol. 42, pp. 76–92, 2017.

[179] G. Pratx and L. Xing, “GPU computing in medical physics: A review,” *Med. Phys.*, vol. 38, no. 5, pp. 2685–2697, 2011.

[180] A. Rajkomar *et al.*, “Scalable and accurate deep learning with electronic health records,” *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1–10, 2018.

[181] B. Tang, Z. Pan, K. Yin, and A. Khateeb, “Recent advances of deep learning in bioinformatics and computational biology,” *Front. Genet.*, vol. 10, p. 214, 2019.

[182] M. Längkvist, L. Karlsson, and A. Loutfi, “A review of unsupervised feature learning and deep learning for time-series modeling,” *Pattern Recognit. Lett.*, vol. 42, pp. 11–24, 2014.

[183] E. Pons, L. M. M. Braun, M. G. M. Hunink, and J. A. Kors, “Natural language processing in radiology: a systematic review,” *Radiology*, vol. 279, no. 2, pp. 329–343, 2016.

[184] P. Perner, “Decision tree induction methods and their application to big data,” in *Modeling and processing for next-generation big-data technologies*, Springer, 2015, pp. 57–88.

[185] D. Goswami, S. Kalkan, and N. Krüger, “Bayesian classification of image structures,” in *Scandinavian Conference on Image Analysis*, 2009, pp. 676–685.

[186] S. Raudys, *Statistical and Neural Classifiers: An integrated approach to design*. Springer Science & Business Media, 2012.

[187] Y. Ma and G. Guo, *Support vector machines applications*, vol. 649. Springer, 2014.

[188] D. Wettschereck, D. W. Aha, and T. Mohri, “A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms,” *Artif. Intell. Rev.*, vol. 11, no. 1, pp. 273–314, 1997.

[189] L. Lu, Y. Zheng, G. Carneiro, and L. Yang, “Deep learning and convolutional neural networks for medical image computing,” *Adv. Comput. Vis. Pattern Recognit.*, vol. 10, pp. 973–978, 2017.

[190] A. K. H. Tung, "Rule-based Classification," in *Encyclopedia of Database Systems*, SpringerLink, 2009, pp. 109-154, DOI: https://doi.org/10.1007/978-0-387-39940-9_559

3. The State-of-the Art in Bioinformatics

Bioinformatics is an interdisciplinary field that deploys methods and software tools for a better understanding of biological data, especially when the data sets are large and complex. Bioinformatics combines biology, information engineering, computer science, mathematics, and statistics to analyse and interpret the biological data. Some common uses of bioinformatics involve the identification of candidate genes and single nucleotide poly-morphisms (SNPs), with the aim of better understanding the genetic basis of a disease, unique adaptations and differences between populations. In a lesser extent, bioinformatics also tries to comprehend the organizational principles within nucleic acid and protein sequences, named proteomics.

This overview presents the most current methods and platforms that are used for sequencing, also referred as Next Generation Sequencing, the genome editing including the most current state-of-the-art technologies, called CRISPR and topics related with Translational Bioinformatics (TBI), which is focused on the convergence of bioinformatics with clinical healthcare. Finally, a summary of the national genomic initiatives is presented, and the ongoing challenges and the future landscape of bioinformatics are discussed.

3.1 Methods in Next-Generation Sequencing

Next generation sequencing (NGS), or massively parallel sequencing, refers to any DNA high-throughput sequencing technology which has revolutionised genomic research, in the sense that an entire human genome can be sequenced within a single day [1]. A typical workflow of an NGS platform includes four steps: 1) Library preparation, 2) Sequencing, 3) Reconstruction and 4) Data Analysis.

3.1.1 Library Preparation

Library preparation is an important procedure for the success of the NGS workflow. This step involves the preparation of DNA or RNA samples to be compatible with a sequencer. The sequencing libraries are generated by fragmenting DNA, adding specialized adapters to both ends. Depending on the method applied, these adapters usually offer complementary strands that allow the DNA fragments to bind to the flow cell. To be more resource efficient, multiple libraries can be combined and sequenced in the same run—a process known as multiplexing. During adapter ligation, unique index sequences are added to each library. These unique sequences are used to discriminate between the libraries during data analysis. Although it depends on application (e.g., gene expression or DNA methylation analysis), the fundamental steps of library preparation include:

1. DNA fragmentation/target selection. In order to separate DNA into smaller pieces, the DNA may be fragmented using physical (e.g., sonication [2] or hydrodynamic shearing [3]), enzymatic (e.g., transposase [4], DNase I [5], or any other restriction endonuclease [6]) or chemical methods (such as heating and the divalent metal cation method [7]). These libraries are referred as fragment libraries.
2. Adapter sequences are annealed to the 5’ and the 3’ end of the fragmented or amplicon DNA. There are two different adapter sequences that can anneal to the DNA fragments and either the 5’ or 3’

orientation. One adapter sequence contains the primer annealing site for the sequencing primer and the second adapter sequence is used to anchor the DNA fragment to a surface for sequencing.

3. The appropriate size selection is needed for the sequencing run. There are two commonly used size selection methods. The first one is the gel electrophoresis method [8], where adapter library fragments are run on a gel to separate the fragments by size. The band which corresponds to the size of interest is collected. The second method used is the bead-based method [9], where magnetic beads are utilized with different concentrations to isolate the DNA fragment sizes of interest.
4. Library quantification which refers to various methods for determining the number of nucleic acid molecules present in the given library. The most common methods include qPCR, fluorimetry, Spectrophotometry, and electrophoretic methods [10].

3.1.2 Sequencing

Generally, the methods of sequencing are classified as short-read (SRS) and long-read (LRS) sequencing. Using next-generation SRS, DNA is divided into short fragments that are amplified (copied) and then sequenced to produce 'reads'. Bioinformatic techniques are applied to piece together the reads into a continuous genomic sequence. The typical range of SRS systems is 75-500 base pairs (bp). On the other hand, as the name implies, long-read sequencing refers to methods capable of sequencing longer strands of DNA by reading single DNA molecules. The typical read lengths for LRS are 10,000 - 100,000bp, while some LRS platforms have reported to produce sequence reads of 882,000bp [11] and other of over 2,000,000bp [12].

3.1.2.1 Short-read sequencing

According to [13], short-read sequencing approaches are divided into two categories: sequencing by ligation (SBL) and sequencing by synthesis (SBS).

1. Sequencing by ligation: SBL approaches include the hybridization and ligation of labelled probe and anchor sequences to a DNA strand. The probes encode one or two known bases and a series of degenerate or universal bases, driving complementary binding between the probe and template, whereas the anchor fragment encodes a known sequence that is complementary to an adapter sequence and provides a site to initiate ligation. After ligation, the template is imaged and the known base or bases in the probe are identified. A new cycle begins after complete removal of the anchor–probe complex or through cleavage to remove the fluorophore and to regenerate the ligation site. During the cycles, single-nucleotide offsets are introduced to ensure every base in the template strand is sequenced. Commercial platforms such as SOLiD by ThermoFisher and DNBSEQ-G400 by MGI (a subsidiary of BGI Group) perform short-read DNA sequencing.

2. Sequencing by synthesis: SBS describes numerous DNA-polymerase-dependent methods. These approaches can be classified either as cyclic reversible termination (CRT) or as single-nucleotide addition (SNA) - also called pyrosequencing.

(a) CRT approaches are defined by their use of terminator molecules that are like those used in Sanger sequencing, in which the ribose 3' - OH group is blocked, preventing elongation [14]. In the beginning of the process, a DNA template is primed by a sequence that is complementary to an adapter region and initiates polymerase binding to this double-stranded DNA (dsDNA) region. During each cycle, a mixture of all four individually labelled and 3' -blocked deoxynucleotides (dNTPs) are added. After the integration of a single dNTP to each elongating complementary strand, unbound dNTPs are removed, and the surface is

imaged to identify which dNTP was incorporated at each cluster. Then, the fluorophore and blocking group can be removed and a new cycle can start. Currently, the Illumina CRT system accounts for the largest market share for sequencing instruments in comparison with other platforms. Illumina's suite of instruments for short-read sequencing varies from small, low-throughput benchtop units to large ultra-high throughput instruments dedicated to population-level whole-genome sequencing (WGS). dNTP identification is achieved through total internal reflection fluorescence (TIRF) microscopy using either two or four laser channels. In most Illumina platforms, each dNTP is bound to a single fluorophore that is specific to that base type and requires four different imaging channels, whereas the NextSeq 550 and Mini-Seq systems use a two-fluorophore system [15].

(b) SNA/Pyrosequencing: SNA approaches rely on a single signal to mark the incorporation of a dNTP into an elongating strand. Thus, each of the four nucleotides must be added iteratively to a sequencing reaction to ensure that only one dNTP is responsible for the signal. Furthermore, this does not require the dNTPs to be blocked, as the absence of the next nucleotide in the sequencing reaction prevents elongation. In the early stage of pyrosequencing [16], as a dNTP is incorporated into a strand, an enzymatic cascade occurs, resulting in a bioluminescence signal. Each burst of light, detected by a charge-coupled device (CCD) camera, can be attributed to the incorporation of one or more identical dNTPs at a particular bead. However, platforms such as the Ion Torrent, instead of using an enzymatic cascade to generate a signal, they detect the H^+ ions that are released as each dNTP is incorporated. The change in pH is detected by an integrated complementary metal-oxide-semiconductor (CMOS) and an ion-sensitive field-effect transistor (ISFET) [17].

3.1.2.2 Long-read sequencing

Long-read sequencing offers reads of more than several kilobases, allowing the resolution of large structural features. Such long reads can connect complex or repetitive regions with a single continuous read, making clearer the positions or size of genomic elements. There are two dominant types of long-read technologies, which refer to single-molecule long-read sequencing approaches, and synthetic approaches which depend on existing short read technologies to construct long reads in silico.

1. Single-molecule long-read sequencing: The single-molecule approaches differ from short-read approaches in that they neither rely on a clonal population of amplified DNA fragments to generate detectable signal, nor do they require chemical cycling for each dNTP added. The most widely long-read platforms are the single-molecule real-time (SMRT) sequencing approach used by Pacific Biosciences (PacBio) and the nanopore sequencing approach from Oxford Nanopore Technologies (ONT).

(a) SMRT/PacBio: The technology used by PacBio is based on the natural process that occurs with the division of cells. Prior to division, DNA is replicated by enzymes called DNA polymerases which efficiently duplicate entire genomes by reading the DNA and sequentially building a complementary strand with matching nucleotides. PacBio utilizes the power of polymerase as a sequencing engine by 'eavesdropping' on it, while it works to replicate DNA. This approach is enabled by two technologies. The first is the phosphor-linked nucleotides to visualize polymerase activity. In contrast to other sequencing approaches, in SMRT phospho-linked nucleotides carry their fluorescent label on the terminal phosphate rather than the base. Through this innovation the enzyme cleaves away the fluorescent label as part of the incorporation process leaving behind a completely natural strand of the DNA. This enables to exploit the inherent properties of the DNA polymerase, including high speed, long read length and high fidelity. The second technology refers to a nanophotonic visualization chamber and is called the zero-mode wave guide (ZMW). The ZMW is a cylindrical metallic chamber approximately 70 nm wide that is illuminated through its glass

support creating an extremely small detection volume of 20 zeptoliter. ZMW technology enables the observation of the individual molecules against the required background of labelled nucleotides, maintaining the same time high signal-to-noise ratio. Nucleotides diffuse in and out of the ZMW in microseconds. When the polymerase meets the correct nucleotide, it takes several milliseconds to incorporate it, during which time its fluorescent label is excited emitting light that is captured by a sensitive detector. After incorporation the fluorescent label is cleaved off and diffuses away. The whole process repeats, generating sequential bursts of light corresponding to the different nucleotides. These are recorded by building the DNA sequence and reading it at a rate of 10bp per second [18]. SMRT by PacBio provides two different sequencing modes: 1) the Continuous Long Read (CLR) sequencing and 2) the Circular Consensus Sequencing (CCS). CLR sequencing is preferred to make long reads of > 50kb, though with a reduced accuracy of (75–90%). On the other hand, CCS enables high read lengths of 10-20 kb with average sequence identity greater than 99% from a single molecule, also known as single molecule high-fidelity HiFi sequencing.

(b) Nanopore sequencing/ONT: In 2014, MinION - the first consumer product of ONT — became commercially available. In comparison with other platforms, nanopore sequencers do not monitor incorporations or hybridizations of nucleotides guided by a template DNA strand. This technology is based on a nanopore that is inserted into an electrically resistant membrane created from synthetic polymers. A potential is applied across the membrane resulting in a current flowing through the aperture at the nanopore. ONT utilizes a strand sequencing method in which intact DNA strands are processed by the nanopores and analyzed in real time. The DNA strands to be sequenced are mixed with copies of a specific enzyme and as the DNA enzyme complex approaches the nanopore, the DNA is pulled through the aperture of the nanopore. The enzyme binds to a single-stranded leader at the end of the dsDNA template and unzips the double strand, feeding it to the nanopore. As the DNA moves through the pore, different k-mer combinations of nucleotides create a characteristic disruption in the electrical current. The instrument has more than 1.000 signals, one for each possible k-mer, especially when modified bases present on native DNA are taken into consideration [19]. Furthermore, the speed of the enzyme can be controlled (e.g., via temperature regulation). The faster it runs the more data is yielded per second, although there is a greater possibility for detection loss. By preparing the DNA to have a hairpin structure at the opposite end, the system can read both strands of the double stranded DNA in one continuous read. PromethION, the latest instrument of ONT, includes 48 flow cells per device, each with 3000 pores (meaning 144,000 pores in total), with each pore running at 500bp/second.

2. Synthetic long-reads: The synthetic approaches do not generate actual long-reads, but they are closer to library preparation approach that leverages ''barcodes'' to allow computational assembly of a larger fragment. These approaches separate large DNA fragments into either microtitre wells or an emulsion such that very few molecules exist in each partition. Within each partition the template fragments are sheared and barcoded. Then, the fragments are sequenced on existing short-read instrumentation, in which data are split by barcode and reassembled with the knowledge that fragments sharing barcodes are derived from the same original large fragment. Synthetic barcoded reads provide an association among small fragments extracted from a larger one. By separating the fragments, repetitive or complicated regions can be isolated, allowing each to be assembled locally. This prevents unresolvable branch points in the assemblies, which lead to breaks (gaps) and shorter assembled contiguous sequences. Moleculo Inc. which was acquired by Illumina in 2013 and 10X Genomics, both utilize platforms that generate synthetic long reads.

3.1.3 Reconstruction

As mentioned above, DNA sequencing is performed on fragments of the DNA strand. After the nucleotide identification of the fragments, these must be recombined in the 'correct' order to match the original DNA strand. From the use of amplification methods many DNA fragments are produced, thus the core idea to solve the reconstruction problem is to use overlapping reads of the fragments. The early algorithms for the reconstruction task were the so-called greedy algorithms (i.e., algorithms that make a sequence of choices, each choice being in some way the best available at that time [20]), which align the DNA fragments based on a similarity score [21]. Despite their simplicity, greedy algorithms do not provide accurate results for large scale combinatorial problems, and other problems that arise due to the sequence procedure (e.g., erroneous reads) and the intrinsic nature of the DNA (e.g., repetitive similar DNA motifs in a single DNA strand) make these algorithms inefficient and only applicable to certain cases. Other algorithms that have been proposed are the following:

1. Hamiltonian path/TSP: The string reconstruction problem can be approached by defining a directed graph G1, where every occurrence of a k-mer in the spectrum is represented by a node in the graph. Every pair of nodes x, y belonging to the whole set k-mers, is connected by a directed edge e from x to y if the $k-1$ suffix of x is identical to the $k-1$ prefix of y. For instance, {GTC, TCC} are 3-mers being connected by a directed edge, since the 2-mer suffix of GTC equals the 2-mer prefix of TCC. Joining the two k-mers into a sequence relates to a cost. The cost of joining two k-mers is equal to k minus the number of nucleotides that overlap in these k-mers. For instance, two k-mers CCATC and TCTAG may overlap on two nucleotides and create a longer sequence CCATCTAG. Consequently, a cost of joining them is equal to 3. The goal is to visit every node in G1 exactly only once and return to the starting point in such a way that a sum of costs of traversed edges included in the G1 cycle is at its minimum, reducing the string reconstruction problem to the known Hamiltonian path or Traveling Salesman Problem (TSP) [22]. However, this problem is known to be NP-hard (i.e., its computational cost runs exponentially as the size of the problem increases), thus unlikely to admit a polynomial–time algorithm.

2. Eulerian path approach: Pevzner et al. [23] proposed another approach, which reduces the reconstruction problem to the well-known Eulerian path problem, which admits a simple linear-time algorithm. The idea is to construct a graph G2 (Pevzner's graph) whose edges (and not the nodes as in the Hamiltonian path problem) correspond to k-mers, and to find a path in the graph that visits every edge only once. Here the nodes are the full set of $(k-1)-mer$ appearing in the spectrum. Based on the defined graph G2, the problem is translated in finding a path that visits all edges on G2. The solution is not necessarily unique because it is possible to detect a Eulerian cycle, which creates multiple ambiguous solutions. Multiple (alternative) solutions are manifested as branches in the graph, and unless the number of branches is very small, there is no good way to determine the correct sequence.

3. De Bruijn graph: The de Bruijn graph [24] is constructed easily by extending the idea proposed by Pevzner, by merging all identically labelled nodes into a single node, without changing the number of in/out edges. Then a solution to this problem leads to the Eulerian path approach, although with less ambiguity due to the problem of different root branches. However, again there may be multiple Eulerian paths in the Bruijin graph. A more efficient algorithm that was proposed to tackle this problem, is the paired information Bruijn graph [25]. This algorithm is based on the paired-read sequencing technology, where pair of reads are generated in both ends of each fragment of the genome. So now instead of talking about reads individually, we are talking about pair of reads, separated by a distance called insert size. Thus, given two $k-mers$, if they are apart at a fixed distance in the genome, they are called a paired $k-mer$. The problem of constructing the genome is not solved by $k-mer$ composition, but by paired

$k-mer$ composition. Thus, in the paired de Bruijn graph every paired $k-mer$ is demonstrated as an edge between its paired prefix and paired suffix. Again, all identical nodes are merged, and the genome reconstruction follows the Eulerian path. Note that what differs from the traditional de Bruijn graph is that in the new form, nodes are labelled by pair of $(k-1)-mers$, while in the traditional form, nodes are labelled by just individual $(k-1)-mers$. Due to this difference in the labelling, the paired de Bruijn graph is obtained by fewer merging, thus making it simpler.

3.1.4 Data Analysis

In bioinformatics, genomic data may have different types of DNA, RNA, proteins, and epigenetic marks. The goal of data analysis is usually to improve identification of differentially expressed genes, disease associated single nucleotide polymorphisms (SNPs) or differentially methylated site. As stated in [26], in data analytics it is possible to integrate the same type of genomic data across multiple studies (horizontal integration) or integrate different types of genomic data in the same set of samples (vertical integration). Here only horizontal data analytics are discussed, in terms of the following steps:

1. Data collection and pre-processing: Firstly, a systematic search is executed to determine inclusion/exclusion criteria for identifying, annotating, and preparing datasets for meta-analysis. This process includes special data management consideration and pre-processing protocol.

2. Statistical methods for meta-analysis: A variety of traditional meta-analysis methods have been applied to genomic applications. There are two main categories: combine p-values and combine effect sizes. The first category refers to Fisher's method, Stouffer's method, and modified versions of them. The second category includes fixed, random, or mixed effects models [27]. A comprehensive review for meta-analysis methods can be found in [28].

3. Targeted biological objectives and hypothesis setting: An important prerequisite behind genomic meta-analysis is to identify the targeted biological objective and the underlying hypothesis setting. Tseng et al. [29] demonstrated two hypotheses settings (HSA and HSB) to detect biomarkers differentially expressed (or SNPs associated to disease) in "all studies" or "one or more studies", respectively. Although HSA is more often the desired biological objective, HSB can be considered when study heterogeneity is expected and of research interest (e.g., when studies utilize different tissues). These two hypothesis settings are narrowly related to traditional union-intersection test (UIT) [30] and intersection-union test (IUT) [31]. Later, Song & Tseng [32] proposed a general class of order statistics of p-values and discussed a robust hypothesis setting to relax HSA from the stringent requirement of differential expression in "all studies" to "most studies". A comparative study between different methods and hypothesis settings for transcriptomic meta-analysis is done in [33].

4. Cross-study heterogeneity: Although the major aim of meta-analysis is to enhance statistical power by combining consensus information, the heterogeneities across studies are also often of importance. Heterogeneity is often identified in genomic studies due to different cohorts, experimental protocols, platforms, or tissues that are utilized to generate the data. In the HSB (IUT) hypothesis setting, adaptively weighted concept (a.k.a. subset-based approach) and meta-Lasso approach have been deployed to detect gene-specific subset of studies that contain differential expression [34] or disease association information [35].

3.2 Genome Editing with CRISPR[3]

Genome editing is the use of various technologies to make permanent changes in the genomic DNA sequence of a cell or organism. Early methods used zinc finger nucleases (ZFNs) or transcription activator-like effector nucleases (TALENs), which are expensive, slow, and difficult to implement in comparison to most current state-of-the-art technologies based on CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats).

CRISPR genome editing is based on a natural immune process used by bacteria to defend themselves against invading viruses. CRISPR systems can recognize and cleave complementary DNA sequences, allowing bacteria to remember and destroy viral invaders. In 2013, researchers demonstrated that they could adapt CRISPR/Cas-dependent genome editing for use in mammalian cells [36]. Cas enzymes are part of a bacterial immune system that incorporates short, viral DNA sequences into the bacterial genome. This is a complicated process that is not entirely understood [37]. What is well characterized is that these viral sequences are found at regular intervals, short distances from one to each other in the bacterial genome. The bacterial DNA between these sequences has palindromic repeating patterns, hence the name, clustered regularly inter-spaced short palindromic repeats. The incorporated viral DNA sequences can be translated into guide RNA (gRNA) when needed—that is, if the same kind of virus tries to infect the bacterium again, the CRISPR system can cut the invading viral DNA through use of the gRNA and Cas enzyme. This last step of the bacterial immune process, when the gRNA is combined with Cas and cleaves the target DNA, is what has been adopted for genome editing in laboratories.

The most common applications of CRISPR include:

• *Screening*: Screening identifies a small number of genes (out of the whole genome) involved in a specific physiological effect. Most CRISPR screening is done in cell culture, although some methods have been devised for use in animal models. In CRISPR screening, scientists usually knock out every gene in the genome that could be important, knocking out only one gene per cell. During this process, some cells die, but others survive and become the predominant cell types. Then, the scientists do next generation sequencing (NGS) on the surviving cells to find out which sequences are still present.

• *Gene silencing*: Almost immediately after the discovery of CRISPR genome editing, some researchers produced a mutated Cas9 that could not cut DNA. This catalytically inactive enzyme, dCas9 (dead Cas9), could still be targeted to a specific genomic site. Interestingly, simply by binding to a target site, dCas9 was able to inactivate gene transcription at that site by preventing binding of the cellular transcription machinery to the gene [38]. Without transcription, the relevant protein is not produced, and the gene is effectively silenced until the cell naturally eliminates the dCas9 enzyme. This approach, whereby CRISPR is used to temporarily silence a gene without cutting the DNA, was named CRISPR interference (CRISPRi).

• Gene activation: CRISPRa is a category of methods that use the same dead Cas9 mutant (dCas9) as CRISPRi. However, in CRISPRa, the RNP (ribonucleoprotein, i.e., a complex of Cas enzyme and guide RNA) is used to carry transcriptional activators, which overcome the transcription-blocking effect caused by dCas9, and turn on transcription at target promoter regions within the target gene. Some CRISPRa systems use activating proteins connected to the gRNA itself rather than to the Cas enzyme [39].

• Nuclear organization and epigenetic modifications: You can use fluorescently labelled CRISPR components to help study nuclear organization, enabling easy visualization of target genes within the nucleus. Numerous systems using modifications of CRISPR components have been designed for this purpose [40]. In the nuclei of cells, DNA is usually tightly wound around histone proteins. The term “epigenetics” refers to posttranslational modifications of these histone proteins, as well as methylation of DNA itself. These

[3] This section is based on the content of “The CRISPR basics handbook.”, © 2020 Integrated DNA Technologies.

modifications affect nuclear and DNA behaviour and are regulated by many enzymes. CRISPR technology can be used to direct such enzymes to particular sites in the genome to regulate epigenetic modifications [41].

3.3 Translational Bioinformatics

Translational bioinformatics (TBI) is a multi-disciplinary and rapidly emerging field that involves the deployment of technologies that translate basic molecular, cellular, genetic, and clinical data into knowledge and medical tools. TBI applies novel methods to the storage, analysis, and interpretation of a massive volume of genetics, genomics, multi-omics, and clinical data, including diagnoses, medications, laboratory measurements, imaging, and clinical notes. TBI links the gap between experimental research and real-world applications to human health.

3.3.1 Genomic Data Resources

Reference datasets enable comparative analyses with parallel data from disease-centric studies to determine variants and processes that are associated to disease. This results in a better understanding of the fundamental mechanisms that occur through a plethora of diseases, an improved ability to predict the treatments that work best for specific patients and improved approaches such as genome-based strategies for the early detection, diagnosis, and treatment of the disease.

A variety of publicly available genomic data deposited in different databases are depicted in Table 4 [26]. Among the biggest biotechnology information centers are the European Bioinformatics Institute (EMBL-EBI) and the National Center for Biotechnology Information (NCBI). The results of Genome-Wide Association Studies (GWAS), including DNA genotype and phenotype data, usually populate dbGAP, which is hosted by NCBI. Since genotyping information can theoretically identify the patient ID, a secure access application through dbGAP is necessary to protect the patients' privacy. Gene expression and epigenetic data are often deposited in NCBI GEO (Gene Expression Omnibus) or ArrayExpress. The NCBI SRA (Sequence Read Archive) is a central location for storing sequencing data. The SRA Toolkit provides easy solutions for downloading large files of sequencing data. In addition, there are also growing data resources from large consortium projects. The Cancer Genome Atlas (TCGA) allows users to download open access data, including de-identified data of clinical and demographic features, mRNA or microRNA expression, copy number alterations, protein or phosphoprotein abundance and DNA methylation. Another large consortium that gathers genomic data from different types of cancers is the International Cancer Genome Consortium (ICGC). Moreover, some other well-known genomic data resources are the Roadmap Epigenomics Project, which focuses on genome-wide epigenetic marks; Genotype-Tissue Expression project (GTEx), which produces RNA-seq data from different human tissues; ENCODE (Encyclopedia Of DNA Elements) project, which intends to study the function of genes and the elements that regulate genes throughout the genome. Many datasets are publicly available but not well-annotated; that makes them difficult for use. Databases with standardized uploading protocols are typically easier to use. Finally, sequencing, or genotyping data of human samples often involve specific issues as far as privacy and legal consent are concerned, and hence their datasets need protection through standardized protocols.

Table 3. Genomic data resources.

Resource	URL	Description
dbGAP	www.ncbi.nlm.nih.gov/gap	DNA genotype and phenotype data
ArrayExpress	www.ebi.ac.uk/arrayexpress	Gene expression and epigenetic marks
GEO	www.ncbi.nlm.nih.gov/geo	Gene expression and epigenetic marks
SRA	www.ncbi.nlm.nih.gov/sra	Sequencing data
TCGA	tcga-data.nci.nih.gov/tcga	Multiple types of open access genomic data
ICGC	https://dcc.icgc.org/	Multiple types of genomic data
Roadmap	www.roadmapepigenomics.org	Epigenomics data
GTEx	www.gtexportal.org/home	RNA-seq from different tissues
ENCODE	www.encodeproject.org	Epigenetic and gene expression data

3.3.2 Genomic Annotation Databases

A DNA sequence has much more value if it is possible to annotate the different features like promoters, exons, introns, transposons, etc. The annotation of those regions in a sequence is called structural annotation and is usually accompanied by a further functional annotation that will demonstrate the functions for these different regions.

The most significant annotation for most genomic studies is the reference genome, with the most current release of human reference genome to be GRCh38.p13 (released on 2019/02/28 by Genome Reference Consortium). Reference genomes can be accessed online at Ensembl or the UCSC Genome Browser, among other online locations. At the DNA level, NCBI dbSNP (Single Nucleotide Polymorphism Database) offers a concise annotation of known SNPs, extracted by various sequencing/genotyping projects, such as the 1000 Genomes Project. Regarding the gene structure, Ensembl's Genebuild workflow automatically annotates genes based on existing evidence of mRNA and proteins in public databases [42]. The GENCODE annotation relates the automatic annotation from Ensembl and manual annotation from the HAVANA (Human and Vertebrate Analysis and Annotation) group [43]. Furthermore, Gene Ontology (GO) database provides ontology terms for gene functions in three categories: biological process, molecular function, and cellular component [44]. There are also many databases for pathway annotations, such as KEGG, PathBank, Reactome, WikiPathways, Pathway Commons and BioCyc. Table 5 summarizes the abovementioned annotation databases. There are also some other useful databases to systematically catalog existing biological findings, such as GWAS Catalog (disease association findings), COSMIC (mutations and gene translocation), miRanda (miRNA target genes) and Genomics of Drug Sensitivity in Cancer (GDSC, for drug response in cancer).

Table 4. Annotation databases.

Category	Database	URL
Genome browser	Ensembl	www.ensembl.org/index.html
	UCSC genome browser	genome.ucsc.edu
SNP/indels	dbSNP	www.ncbi.nlm.nih.gov/SNP
Gene structure	GENCODE	www.gencodegenes.org
	Ensembl's Genebuild	www.ensembl.org/index.html
Functional annotation	Pathway Commons	www.pathwaycommons.org
	KEGG	www.genome.jp/kegg
	Gene Ontology (GO)	geneontology.org
	PathBank	pathbank.org/
	Reactome	https://reactome.org/
	WikiPathways	www.wikipathways.org
	BioCyc	https://biocyc.org/

3.3.3 Functional and Clinical Interpretation

When identifying gene mutations, the first step for cancer disease is to catalogue the differences between the healthy and tumor genomes and for other rare diseases to catalogue all the nucleotide differences or variations in a patient's genome compared to a reference genome [45]. The next step is to comprehend the clinical significance of the variants, their inheritance patterns, and the strength of their association to the disease or phenotype. Understanding the clinical and functional significance of each variant demands complex bioinformatics analysis and the integration of numerous other data types:

1. Gene structure information data is necessary to identify if the variant lies in the coding or non-coding region of the genome.

2. Functional data, and data for coding variant and protein structure are necessary to detect the impact of the mutation on protein function.

3. Transcriptomics and proteomics data are required to identify tissue and cell expression profiles.

4. Mutation experimental data from human cell or model organisms and disease variation information are needed to understand associated phenotypes.

5. Biological pathway knowledge and protein interaction network are required to learn more about the function and interaction with other proteins.

6. Data from clinical trials and pharmaceutical agents are also important to identify which medicines target specific proteins or biological pathways. If it is available, longitudinal phenotypic information at the individual and population levels is also needed.

3.3.4 Clinical Data Environment

The incorporation of EHRs (Electronic Health Records) in patient healthcare is very important to relate molecular and clinical data. The implementation of EHRs within hospitals has driven to greater standardization and efficiency [46]. A complete overview of EHRs implementation can be found in [47]. However, one critical point of data integration in EHRs is the appropriate annotation and mapping of data to curated vocabularies or ontologies. For genomic data integration with clinical information, data from primary care, hospitals, outcomes, registries, and social care records should be recorded first, by means of controlled clinical terminologies, such as SNOMED Clinical Terms and the Human Phenotype Ontology [48]. Ontologies as the above-mentioned are never complete, and users such as clinicians will need to work with ontology developers to continuously enhance the precision and accuracy of terminologies. Furthermore, organizations such as CDISC create standards to support the acquisition, sharing, submission and archiving of clinical research data.

Clinical data is usually generated and held across a wide variety of point of care settings such as acute hospitals, general practitioners, community hospitals, mental health, and social care. Thus, to minimize duplication rates, the integration of health data from different sources should populate a common repository.

3.3.5 Data Interoperability

New models for data interoperability have been created so that experts worldwide can access, use, and deposit their data. A pilot project in rare disease, the Deciphering Developmental Disorders (DDD) study, aimed to determine the feasibility of translating new high-throughput genomic technologies into clinical practice, elucidating the underlying genetic architecture of developmental disorders. This study utilized the whole exome sequencing to diagnose 27% of 1,133 previously investigated yet undiagnosed children with developmental disorders [49] and established a unique database model in the DECIPHER database.

DECIPHER database, an international community of academic departments of clinical genetics and rare disease genomics which now numbers more than 250 centers, has uploaded more than 18,000 cases. Each center maintains control of its own patient's data (which are password protected within the center's own DECIPHER project) until consent is given in order to share the data within specific parties in a collaborative group or to allow anonymous genomic and phenotypic data to become freely available within genome browsers. Once data are uploaded and shared, consortium members gain access to the patient report and contact with other partners to discuss patients of interest. After data analysis, pertinent genomic variants are returned to individual research participants through their local clinical genetics team.

Furthermore, the Global Alliance for Global Health (GA4GH) established in 2013, deployed a common framework of approaches for adoption, aiming to accelerate progress in human health, drive efficiencies and decrease costs. The goal of GA4GH is to create a system of servers, generate standard markup languages and develop resources and applications like the implementation of the World Wide Web for users to access genomics information [50]. GA4GH includes institutions like EMBL-EBI that specialized in data management and analysis of big data projects.

3.3.6 Use of Genomic Data and Electronic Health Records

Processed data extracted by genome sequencing projects can also be integrated into existing data derived from other medical systems in a manner that enables precision medicine. Data that can be integrated include:

1. Molecular profiles that distinguish differences between diseased and normal states or provide subclassification of a disease.

2. Annotation of variants with clinical importance in different diseases.

3. Annotation of variants and genes, as well as their interaction with drugs.

4. Biomarkers used for diagnosis and disease monitoring.

5. Reference images that can associate molecular data with disease phenotypes.

6. Human pathogen data and their virulence components.

The use of the abovementioned data will enable mainly the development of specific target disease drugs. One of the major reasons for the high rate of attrition in late-stage clinical trials is the lack of drug efficacy. Often the incorrect gene or protein is chosen as the drug target in early drug development stage, where the premise is that perturbation of this protein by a compound will significantly change the course of disease [51]. Also, in [52] it is shown that genetic data that relate a drug target to a phenotype or disease have higher success rates in the clinic. A detailed review on the availability of public data and the analytical tools used for various data types for target selection and drug discovery can be found in [53].

Reported studies in the literature have successfully stratified patients and detected potential biomarkers of drug response, by utilizing biobanks and integrating different data types. E.g., in a study by Folkersen et al. [54] applied in Rheumatoid Arthritis (RA), a biobank was used to test the claim that the current state-of-the-art precision medicine will benefit RA patients. In fact, high-throughput RNA sequencing, DNA genotyping, extensive proteomics, and flow cytometry measurements, as well as comprehensive clinical phenotyping, led to the identification of a small set of biomarkers available in peripheral blood that predict clinical response to tumor necrosis factor blockade. In another large cohort study of Bagley et al. [55], there was an integration of data derived from electronic medical systems with disease-associated genetic variants data to study the relationship between disease co-occurrence and commonly shared genetic architectures of disease. The study examined 35 disorders, medical records of over 1.2 million patients, and variants from over 17,000 publications, and determined specific shared genes between disease classes that were not previously considered to be related, such as autoimmune and neuropsychiatric disorders. Furthermore, public-private initiatives such as Open Targets, a collaboration between Biogen, the EMBL-European Bioinformatics Institute, GlaxoSmithKline, and the Wellcome Trust Sanger Institute, provide comprehensive and updated relevant genetics and high-throughput genomics data for drug target selection and validation [56]. Currently, genomic data has also enabled very large-scale projects, such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) study, which is an international collaboration to determine common patterns of mutation in more than 2,600 cancer whole genomes from the International Cancer Genome Consortium [57]. The major aim of PCAWG is the generation of genomic, transcriptomic, and epigenomic changes in 50 different tumor types and/or subtypes [58].

3.4 National Genomic Initiatives

According to a publication [59], released in the American Journal of Human Genetics, members of the Global Alliance for Genomics and Health (GA4GH) review the different approaches being taken around the world to integrate genomics into healthcare and demonstrate a roadmap for sharing strategies, standards and data internationally to accelerate implementation. The authors of the study provide a detailed overview of the national genomics strategy in the UK, the USA, France, and Australia. In addition, they

note that Saudi Arabia, Estonia, Finland, Denmark, Japan, and Qatar are all developing their own national strategies, which range from projects that focus on rare disease and cancer—where genomic data will have the most immediate impact—to projects that plan to roll out sequencing services across the healthy population for research purposes that feed back into healthcare and benefit everyone. The generation of genomic data in the healthcare setting will quickly outpace that in research within the next five years, with 60 million genomes expected to be sequenced by 2025. By 2030, China hopes to reach its goal of adding another 100 million genomes through the Chinese Precision Medicine Initiative. A summary of currently active national government-funded genomic medicine initiatives is presented in table 6.

Table 5: Currently active national government-funded genomic medicine initiatives [59].

Country (Population) Healthcare system	**Initiatives** Years active [Ref]	**Focus areas**	**Summary**
Australia (25,000,000) Public: mixed state and federal	**Australian Genomics** 2014 – 2021 [66]	Infrastructure and clinical cohorts: rare diseases, cancer, infectious diseases.	A collaborative partnership of over 80 institutions, with research driven through 4 programs: National diagnostic and research network; National approach to data federation and analysis; Evaluation, policy, ethics; and Workforce and education.
Brazil (207,000,000) Public	**Brazilian Initiative on Precision Medicine** 2015-2024 [67]	Infrastructure, population-based cohort, rare and common disease cohorts.	Collaboration between five Research Innovation and dissemination centers to develop shared genomic databases compliant with GA4GH standards. Initial focus creation of reference datasets, anticipated to progress to clinical cohorts.
Denmark Faroe Islands (5,700,000) Public	**Genome Denmark** 2012- [68]	Infrastructure, population-based cohort, pathogen project.	Consortium between four universities, two hospitals and two private companies that aims to establish a national platform for sequencing and bioinformatics. Two demonstration projects: Cancer and Pathogens and Danish reference genome.
	FarGenProject 2011-2017 [69]	Infrastructure and population-based cohort.	Establish Faroese reference genome and biobank through sequencing 1,500 individuals; develop local competencies in genome sequencing.
Estonia (1,400,000) Public	**Eesti biopangas** Estonian Genome Project 2000-ongoing [70]	Infrastructure and population-based cohort.	Genomic data from GWAS, whole genome and whole exome sequencing from 52,000 individuals is linked to information from questionnaires, health records and physical examination.
Finland (5,490,000) Public	**Finland Genome Strategy** 2015-2020 [71]	Infrastructure	Development of Finnish national reference database and IT infrastructure to enable data integration between genomic data, metadata, and health records. Legal and ethics framework, workforce development,

			clinical decision-support tools, public engagement, and education.
France (67,000,000) Public	**Plan France Médecine Génomique 2025** France Genomic Medicine Plan 2016-2025 [72]	Infrastructure and clinical cohorts: rare diseases, cancer, diabetes.	Initial focus on sequencing patients with cancer, diabetes, and rare conditions, with the establishment of a hub-and-spoke model of 12 sequencing platforms in the country, and two national centers for genomic expertise and analysis. Expected to be capable of processing the equivalent of 235,000 genomes a year by 2020, and it is anticipated that the program will be expanded to common conditions 2020 onwards.
Japan (123,000,000) Public	**Japan Genomic Medicine Program** 2015- [73]	Infrastructure, clinical and population-based cohorts, drug discovery.	The Japan Genomic Medicine Program is one of the strategic priority areas of the Japan Agency for Medical Research and Development (AMED). Five initiatives underpin the Genomic Medicine Program: Tailor-made medical treatment (mapping disease susceptibility and pharmacogenomics in a cohort of 100,000 patients); Platform for promotion of genome medicine (maximizing the efficiency of clinical and research infrastructure, and undertaking a research program in ethics, legal and societal implications); Integrated database of clinical and genomic information; Platform for genomics-based drug discovery; and the Tohoku medical megabank project (developing a cohort of 150,000 individuals with deep phenotyping and genomic analysis).
Netherlands (17,000,000) Public	**RADICON-NL** 2016-2025 [74]	Rare disease	Trialing new technologies such as rapid whole genome sequencing in small cohorts to determine utility.
	Health-Research Infrastructure 2015- [75]	Infrastructure	Establish single interconnected infrastructure to combine genomic and other health data from multiple sources.
Qatar (2,570,000) Public	**Qatar Genome** 2015- [76]	Infrastructure and population-based cohort.	Creation of large data sets of around 6,000 deeply phenotyped individuals with whole genome sequencing data, available to researchers. Workforce development. Development of a national policy on genomic research. Clinical genome interpretation reporting service to be introduced.
Saudi Arabia (32,280,000) Public	**Saudi Human Genome Program** 2013- [77]	Infrastructure, clinical cohorts (rare and common genetic conditions) and population-based cohorts	Aims to sequence 100,000 individuals and create a national network of 7 sequencing laboratories, catalogue Saudi-specific mutations in known disease genes, catalogue normal genetic variation in the Saudi population, catalogue mutations for recessive and

			common genetic disorders of unknown cause.
Switzerland (8,000,000) Public: mixed federal and canton	**Swiss Personalized Health Network** 2017-2020 [78]	Infrastructure.	Development of a distributed federated network, integrating existing heterogeneous systems at partner institutions. Development of common data standards and semantics. Implementation of a general consent. Definition of a data sharing policy framework.
Turkey (79,000,000) Public	**Türkiye Genom Projesi** Turkish Genome Project 2017-2023 [79]	Infrastructure, clinical cohorts (rare disease, cancer, neurological disease) and population-based cohort.	Establish Turkish reference genome and clinical genomics infrastructure. Plans to sequence 100,000 individuals.
United Kingdom (65,640,000) Public	**Genomics England** 2013-2021 [80]	Infrastructure and clinical cohorts: rare diseases, cancer, infectious diseases.	100,000 Genomes project is driving the establishment of the infrastructure required for the delivery of diagnostic genomics services in England, including a centralized sequencing facility, standardized bioinformatics and analysis pipeline, biorepository, and data center. 13 NHS Genomic Medicine Centers recruiting participants and returning results. Health Education England delivering workforce training.
	Scottish Genomes Partnership 2015- [81]	Infrastructure, clinical cohorts (rare disease and cancer), population-based cohort.	Establishment of two sequencing laboratories, and four clinical genomics centers. Recruiting to 100,000 Genomes Project, and compiling reference genomic data.
	Welsh Genomics for Precision Medicine Strategy 2017- [82]	Infrastructure, clinical cohorts (rare disease and cancer), population-based cohort.	Establishment of Genomic Medicine Centre, recruiting to 100,000 Genomes project, and compiling reference genomic data.
	Northern Ireland Genomic Medicine Centre 2017- [83]	Infrastructure, clinical cohorts (rare disease and cancer), population-based cohort.	Establishment of Northern Ireland Genomic Medicine Centre, recruiting to the 100,000 Genomes project, and compiling reference genomic data.
USA (321,000,000) Mixed: private insurance and public	**National Human Genome Research Institute (NHGRI)** 2007- [84]	Infrastructure and clinical cohorts.	Identification of barriers to implementation of genomics in clinical care and development of solutions and best practices for widespread dissemination. Landmark projects include the Undiagnosed Diseases Network (UDN), Clinical Sequencing Evidence-Generating Research (CSER) consortium, Electronic Medical Records and Genomics (eMERGE) Network, Implementing Genomics in Practice (IGNITE) Network and the Newborn Sequencing in Genomic Medicine

			and Public Health (NSIGHT) program, and the Clinical Genomics (ClinGen) Resource.
	Precision Medicine Initiative (All of Us) 2016-2025 [85]	Population-based cohort.	Aims to create one of the largest, most diverse biomedical datasets through engaging 1,000,000 volunteers and combining genomic data with information from electronic health records, questionnaires, physical evaluations, and biosensors.

3.5 Ongoing Challenges

3.5.1 Standardization

Bioinformatics analysis that leads to clinical interpretation is an expensive part of the workflow, as the storage and computation costs have not been reduced as quickly as the sequencing costs. Currently, the bottlenecks in genomic medicine lie in the data analysis and interpretation end of the pipeline. Interpretation at this case is a crucial step since a patient's diagnosis status and potential treatment options are dependent on interpretation, and not on the raw or processed sequence data [45].

Efforts to determine a gold standard methodology and evaluate the performance of data analytical methods are currently emerging, including a study by Tokheim and colleagues [60], who compared eight different algorithms to identify which gene variants drove cancer driver genes and which were simply passenger mutations. Furthermore, in [61], a methodology is introduced to verify systems biology research workflows that are increasingly complex and sophisticated in industrial and academic settings. This methodology named 'Industrial Methodology for Process Verification in Research', or IMPROVER, is based on the evaluation of a research program by dividing a workflow into smaller building blocks that are individually verified. The verification of each building block can be implemented internally by members of the research program or externally by 'crowd-sourcing' to an interested community (www.sbvimprover.com).

In conjunction with the healthcare system, Electronic Medical Records (EMR) represent an easy source of coded medical data, but the lack of standards and the variation among the different systems can demonstrate inaccuracies and biases when this data is used for analyses such as calculating disease prevalence, incidence, etc. Kevin Wilson et al. [62] review the current methods employed to evaluate diagnostic tests. There is also a further need for standardization around clinical data capture and communication. The challenges lie in being able to gather this information from busy clinicians, and the data also needs to be integrated across the various points of patient care.

Finally, Mark Caulfield, chief scientist at Genomics England pointed that nearly every nation in the world has a different way of delivering healthcare, so no one approach will fit all countries. But for genomic data to fully deliver on its promise, scientists need to share both their expertise and the data to enable enhanced patient care.

3.5.2 Data Storage and Sharing

Currently, a variety of human genomic data generated so far populated the public databases for broad research use [45]. Human genomic and phenotypic data from clinical or research studies which would require a researcher to have a signed agreement with the originating body are largely stored in controlled-

access repositories such as the European Genome-Phenome Archive (EGA), the NIH database of Genotypes and Phenotypes (dbGaP). However, due to the different ethical and legal systems of each country, these systems are not scalable nor suitable for the growing volume of genomic data from national health studies.

Managed storage systems following the national legislation and allowing access to data for research purposes are crucial. Researcher access to genomic databases is needed to create a research community that will be connected and may contribute directly to national health services and patient care systems. Analyses that use the variety of data from hundreds of thousands of patients coming from multiple healthcare systems will add much more to our knowledge of the genetic basis of disease than multiple individual studies using small sample cohorts from individual healthcare systems.

New data sharing mechanisms are also needed to minimize the movement of large volumes of data. Cloud computing frameworks allow remote storage, with analysis scripts uploaded to the cloud and analysis performed remotely on virtual machines physically located at the remote site "next to" the data. This reduces data transfer requirements since only the scripts and analytical results are transferred to and from the analysts' institution or desktop, whereas data populates permanently in the cloud [63]. Initiatives such as the European Open Science Cloud will help further the creation of infrastructures to enable data sharing and service provision across borders and disciplines.

With health data from large numbers of people, it will be critical to find ways to protect individuals' privacy and the confidentiality of their health information, while enabling research to take place at the same time. Current practices for researcher access to data that include paper-based agreements among users, institutions, and data access committees must be replaced by electronic mechanisms, enhancing, and strengthening the connection between basic and clinical research.

3.5.3 Biomedical Informatics Coordination

The authors in [45] propose the development of a 'Biomedical Informatics Institute' to act as a driver and coordinating center for health and biomedical informatics research in each country. This center should act along with existing medical research and informatics organizations to form an integrated network with hospitals, research organizations, and local and international health initiatives to maximize the utility of electronic health data. In bigger nations, this institute would itself likely be a network, but with a center of gravity, or hub, at or within one institute. In that respect, local research bioinformatics institutes will be responsible for handling and providing both public and controlled access data, whereas each national biomedical informatics institute will be responsible for data and services that need to stay within the national framework.

Such centers would be the natural partners for research bioinformatics organizations such as EMBL-EBI or NCBI. In European countries, the development of biomedical informatics institutes or networks may be coordinated through an ELIXIR node: ELIXIR is the European life-science infrastructure for biological data.

3.6 Future Landscape

Over the last few decades, open-source data that permits data reuse and data integration has made possible great progress in molecular biology. These advances range from recombinant DNA drugs, animal cloning and gene therapy. Although genomic data are mainly generated for disease diagnoses, treatment and prevention, the availability of these data for use in research can result in a better understanding of

disease mechanisms and will lead to improvements in treatment strategies. Moreover, the use of bioinformatics in healthcare will further assist to fundamental discoveries related to the big questions of biology.

In order to leverage the growing and already huge amount of data, translational bioinformatics methods and resources will need to evolve to include algorithms for streaming data capture, real-time data aggregation, machine learning, predictive analytics, and visualization solutions to integrate health monitoring data with EMRs and genomics data [64].

Finally, if genomics medicine approaches become part of routine healthcare, doctors and other healthcare providers will require better grounding in molecular genetics and biochemistry. They will need to interpret the results of genetic tests, understand how that information is relevant to treatment or prevention approaches, and convey this knowledge to patients. In addition, education in the data sciences is also very important [65]. Programs to ensure the long-term generation of proficient investigators who understand the multi-disciplinary nature of genomics in clinical practice and research, should be established and may create a new medical discipline.

3.7 References

[1] S. Behjati and P. S. Tarpey, "What is next generation sequencing?," *Arch. Dis. Childhood-Education Pract.*, vol. 98, no. 6, pp. 236–238, 2013.

[2] Q. Tseng, A. M. Lomonosov, E. E. M. Furlong, and C. A. Merten, "Fragmentation of DNA in a submicroliter microfluidic sonication device," *Lab Chip*, vol. 12, no. 22, pp. 4677–4682, 2012.

[3] A. Joneja and X. Huang, "A device for automated hydrodynamic shearing of genomic DNA," *Biotechniques*, vol. 46, no. 7, pp. 553–556, 2009.

[4] A. S. Belyaev, "Immobilized transposase complexes for DNA fragmentation and tagging." Google Patents, May 09, 2017.

[5] A. Hell, G. D. Birnie, T. K. Slimming, and J. Paul, "Controlled fragmentation of DNA by DNase I," *Anal. Biochem.*, vol. 48, no. 2, pp. 369–377, 1972.

[6] R. J. Roberts, "How restriction enzymes became the workhorses of molecular biology," *Proc. Natl. Acad. Sci.*, vol. 102, no. 17, pp. 5905–5908, 2005.

[7] V. W. Campbell and D. A. Jackson, "The effect of divalent cations on the mode of action of DNase I. The initial reaction products produced from covalently closed circular DNA.," *J. Biol. Chem.*, vol. 255, no. 8, pp. 3726–3735, 1980.

[8] P. Y. Lee, J. Costumbrado, C.-Y. Hsu, and Y. H. Kim, "Agarose gel electrophoresis for the separation of DNA fragments," *JoVE (Journal Vis. Exp.*, no. 62, p. e3923, 2012.

[9] C. K. Y. Tan, J. A. Cowley, and D. R. Jerry, "A magnetic bead-based DNA extraction protocol suitable for high-throughput genotyping in shrimp breeding programs," *Genet. Aquat. Org.*, vol. 3, no. 2, pp. 47–56, 2019.

[10] "The Quantitation Question: How does accurate library quantitation influence sequencing?," *BioLabs, New England*. https://international.neb.com/tools-and-resources/feature-articles/the-quantitation-question-how-does-accurate-library-quantitation-influence-sequencing.

[11] M. Jain *et al.*, "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nat. Biotechnol.*, vol. 36, no. 4, pp. 338–345, 2018.

[12] A. Payne, N. Holmes, V. Rakyan, and M. Loose, "BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files," *Bioinformatics*, vol. 35, no. 13, pp. 2193–2198, 2019.

[13] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nat. Rev. Genet.*, vol. 17, no. 6, p. 333, 2016.

[14] J. Guo *et al.*, "Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides," *Proc. Natl. Acad. Sci.*, vol. 105, no. 27, pp. 9145–9150, 2008.

[15] Illumina, "Evolution of 2-Channel SBS Technology." https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html.

[16] M. Margulies *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, pp. 376–380, 2005.

[17] J. M. Rothberg *et al.*, "An integrated semiconductor device enabling non-optical genome sequencing," *Nature*, vol. 475, no. 7356, pp. 348–352, 2011.

[18] S. Ardui, A. Ameur, J. R. Vermeesch, and M. S. Hestand, "Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics," *Nucleic Acids Res.*, vol. 46, no. 5, pp. 2159–2168, 2018.

[19] K. M. A. Gartland, M. Dundar, T. Beccari, M. V. Magni, and J. S. Gartland, "Advances in biotechnology: Genomics and genome editing," *EuroBiotech J.*, vol. 1, no. 1, pp. 2–9, 2017.

[20] S. A. Curtis, "The classification of greedy algorithms," *Sci. Comput. Program.*, vol. 49, no. 1–3, pp. 125–157, 2003.

[21] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences," *J. Comput. Biol.*, vol. 7, no. 1–2, pp. 203–214, 2000.

[22] J. Błażewicz, P. Formanowicz, M. Kasprzak, W. T. Markiewicz, and J. Węglarz, "DNA sequencing with positive and negative errors," *J. Comput. Biol.*, vol. 6, no. 1, pp. 113–123, 1999.

[23] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proc. Natl. Acad. Sci.*, vol. 98, no. 17, pp. 9748–9753, 2001.

[24] P. E. C. Compeau, P. A. Pevzner, and G. Tesler, "How to apply de Bruijn graphs to genome assembly," *Nat. Biotechnol.*, vol. 29, no. 11, pp. 987–991, 2011.

[25] P. Medvedev, S. Pham, M. Chaisson, G. Tesler, and P. Pevzner, "Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers," *J. Comput. Biol.*, vol. 18, no. 11, pp. 1625–1634, 2011.

[26] S. Richardson, G. C. Tseng, and W. Sun, "Statistical methods in integrative genomics," *Annu. Rev. Stat. its Appl.*, vol. 3, pp. 181–209, 2016.

[27] T. A. Trikalinos, G. Salanti, E. Zintzaras, and J. P. A. Ioannidis, "Meta-analysis methods," *Adv. Genet.*, vol. 60, pp. 311–334, 2008.

[28] F. Begum, D. Ghosh, G. C. Tseng, and E. Feingold, "Comprehensive literature review and statistical considerations for GWAS meta-analysis," *Nucleic Acids Res.*, vol. 40, no. 9, pp. 3777–3784, 2012.

[29] J. Li and G. C. Tseng, "An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies," *Ann. Appl. Stat.*, vol. 5, no. 2A, pp. 994–1019,

2011.

[30] F. Pesarin, L. Salmaso, E. Carrozzo, and R. Arboretti, "Union–intersection permutation solution for two-sample equivalence testing," *Stat. Comput.*, vol. 26, no. 3, pp. 693–701, 2016.

[31] R. Arboretti, E. Carrozzo, F. Pesarin, and L. Salmaso, "Testing for equivalence: an intersection-union permutation solution," *Stat. Biopharm. Res.*, vol. 10, no. 2, pp. 130–138, 2018.

[32] C. Song and G. C. Tseng, "Hypothesis setting and order statistic for robust genomic meta-analysis," *Ann. Appl. Stat.*, vol. 8, no. 2, p. 777, 2014.

[33] L.-C. Chang, H.-M. Lin, E. Sibille, and G. C. Tseng, "Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–15, 2013.

[34] Q. Li, S. Wang, C. Huang, M. Yu, and J. Shao, "Meta-analysis based variable selection for gene expression data," *Biometrics*, vol. 70, no. 4, pp. 872–880, 2014.

[35] S. Bhattacharjee *et al.*, "A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits," *Am. J. Hum. Genet.*, vol. 90, no. 5, pp. 821–835, 2012.

[36] E. S. Lander, "The heroes of CRISPR," *Cell*, vol. 164, no. 1–2, pp. 18–28, 2016.

[37] F. Hille and E. Charpentier, "CRISPR-Cas: biology, mechanisms and relevance," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 371, no. 1707, p. 20150496, 2016.

[38] L. S. Qi *et al.*, "Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression," *Cell*, vol. 152, no. 5, pp. 1173–1183, 2013.

[39] S. Konermann *et al.*, "Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex," *Nature*, vol. 517, no. 7536, pp. 583–588, 2015.

[40] Y. Hong, G. Lu, J. Duan, W. Liu, and Y. Zhang, "Comparison and optimization of CRISPR/dCas9/gRNA genome-labeling systems for live cell imaging," *Genome Biol.*, vol. 19, no. 1, pp. 1–10, 2018.

[41] L. R. Polstein and C. A. Gersbach, "A light-inducible CRISPR-Cas9 system for control of endogenous gene activation," *Nat. Chem. Biol.*, vol. 11, no. 3, pp. 198–200, 2015.

[42] V. Curwen *et al.*, "The Ensembl automatic gene annotation system," *Genome Res.*, vol. 14, no. 5, pp. 942–950, 2004.

[43] J. Harrow *et al.*, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Res.*, vol. 22, no. 9, pp. 1760–1774, 2012.

[44] M. Ashburner *et al.*, "Gene ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, 2000.

[45] J. Vamathevan and E. Birney, "A review of recent advances in translational bioinformatics: bridges from biology to medicine," *Yearb. Med. Inform.*, vol. 26, no. 1, p. 178, 2017.

[46] R. J. Johnson, "A comprehensive review of an electronic health record system soon to assume market ascendancy: EPIC," *J Heal. Commun*, vol. 1, no. 4, p. 36, 2016.

[47] L. Nguyen, E. Bellucci, and L. T. Nguyen, "Electronic health records implementation: an evaluation of information system impact and contingency factors," *Int. J. Med. Inform.*, vol. 83, no. 11, pp. 779–796, 2014.

[48] S. Köhler *et al.*, "The human phenotype ontology in 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D865–D876, 2017.

[49] C. F. Wright *et al.*, "Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data," *Lancet*, vol. 385, no. 9975, pp. 1305–1314, 2015.

[50] S. J. Aronson and H. L. Rehm, "Building the foundation for genomics in precision medicine," *Nature*, vol. 526, no. 7573, pp. 336–342, 2015.

[51] M. J. Waring *et al.*, "An analysis of the attrition of drug candidates from four major pharmaceutical companies," *Nat. Rev. Drug Discov.*, vol. 14, no. 7, pp. 475–486, 2015.

[52] M. R. Nelson *et al.*, "The support of human genetic evidence for approved drug indications," *Nat. Genet.*, vol. 47, no. 8, pp. 856–860, 2015.

[53] B. Chen and A. J. Butte, "Leveraging big data to transform target selection and drug discovery," *Clin. Pharmacol. Ther.*, vol. 99, no. 3, pp. 285–297, 2016.

[54] L. Folkersen *et al.*, "Integration of known DNA, RNA and protein biomarkers provides prediction of anti-TNF response in rheumatoid arthritis: results from the COMBINE study," *Mol. Med.*, vol. 22, no. 1, pp. 322–328, 2016.

[55] S. C. Bagley, M. Sirota, R. Chen, A. J. Butte, and R. B. Altman, "Constraints on biological mechanism from disease comorbidity using electronic medical records and database of genetic variants," *PLoS Comput. Biol.*, vol. 12, no. 4, p. e1004885, 2016.

[56] G. Koscielny *et al.*, "Open Targets: a platform for therapeutic target identification and validation," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D985–D994, 2017.

[57] I. The, T. P.-C. A. of Whole, and G. Consortium, "Pan-cancer analysis of whole genomes," *Nature*, vol. 578, no. 7793, p. 82, 2020.

[58] A. Biankin, J. L. Jennings, and L. D. Stein, "International Cancer Genome Consortium." AACR, 2018.

[59] Z. Stark *et al.*, "Integrating genomics into healthcare: a global responsibility," *Am. J. Hum. Genet.*, vol. 104, no. 1, pp. 13–20, 2019.

[60] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin, "Evaluating the evaluation of cancer driver genes," *Proc. Natl. Acad. Sci.*, vol. 113, no. 50, pp. 14330–14335, 2016.

[61] P. Meyer *et al.*, "Industrial methodology for process verification in research (IMPROVER): toward systems biology verification," *Bioinformatics*, vol. 28, no. 9, pp. 1193–1201, 2012.

[62] C. M. Umemneku Chikere, K. Wilson, S. Graziadio, L. Vale, and A. J. Allen, "Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard–An update," *PLoS One*, vol. 14, no. 10, p. e0223832, 2019.

[63] G. A. for G. and Health*, "A federated ecosystem for sharing genomic, clinical data," *Science (80-.).*, vol. 352, no. 6291, pp. 1278–1280, 2016.

[64] K. Shameer, M. A. Badgeley, R. Miotto, B. S. Glicksberg, J. W. Morgan, and J. T. Dudley, "Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams," *Brief. Bioinform.*, vol. 18, no. 1, pp. 105–124, 2017.

[66] Australian Genomics Health Alliance (Australian Genomics), URL: https://www.melbournebioinformatics.org.au/project/austgenomics/

[67] Brazilian Initiative on Precision Medicine, URL: https://bipmed.org/

[68] GenomeDenmark, URL: http://www.genomedenmark.dk/english/

[69] Faroe Genome Project, URL: https://www.fargen.fo/en/about-fargen/project/

[70] A. Metspalu, "The Estonian Genome Project," *Drug Development Research,* June 2004

[71] Finland Genome Strategy, URL: https://julkaisut.valtioneuvosto.fi/handle/10024/74712

[72] PLAN FRANCE MÉDECINE, URL: https://pfmg2025.aviesan.fr/

[73] Genomic Medicine in Japan, URL: https://www.genome.gov/Multimedia/Slides/GM6/21_Okamura_Kubo_Miyano_Japan.pdf

[74] Radicon-NL, URL: https://www.wgs-first.nl/en/project

[75] Health-RI, URL: https://www.dtls.nl/large-scale-research-infrastructures/health-ri/

[76] QATAR GENOME PROGRAMME, URL: https://qatargenome.org.qa/

[77] SAUDI HUMAN GENOME PROGRAM, URL: https://shgp.kacst.edu.sa/index.en.html#home

[78] Swiss Personalized Health Network, URL: https://sphn.ch/

[79] TURKISH GENOME PROJECT, URL: https://www.bbmri-eric.eu/news-events/turkish-genome-project-launched/

[80] Genomic England, URL: https://www.wellcomegenomecampus.org/aboutus/genomics-england/

[81] Scottish Genomes Partnership, URL: https://www.scottishgenomespartnership.org/

[82] Welsh Genomics for Precision Medicine Strategy, URL: https://gov.wales/sites/default/files/publications/2019-04/genomics-for-precision-medicine-strategy.pdf

[83] The Northern Ireland Genomic Medicine Centre, URL: https://gtr.ukri.org/projects?ref=MC_PC_16018

[84] NHGRI, URL: https://www.genome.gov/about-genomics

[85] L. P. PL Sankar, "The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues," *Genetics in Medicine,* 2017

4. The State of the Art in Sensor Informatics

Wearable devices enable the continuous monitoring of different physiological parameters offering innovative solutions like the prevention of diseases and promoting a healthy lifestyle and wellbeing. By ensuring the fidelity of the produced data, wearables can also have an impact on clinical decision-making. Furthermore, with the help of wearable sensors doctors and other actors from the healthcare sector can monitor the progress of a patient without being in proximity, thus enabling personalized patient care and reducing healthcare costs, e.g. shorter rehabilitation periods in hospitals.

4.1 Data Types and Acquisition Techniques

Taking measurements of the four primary vital signs (temperature, heart rate, respiration rate, and blood pressure) is limited by the number of visits to a healthcare actor, while traditional in-patient devices are bulky and not easy portable. Furthermore, the use of secondary data, such as sweat, emotional state etc., cannot be measured by conventional healthcare settings and is often omitted in the decision-making process. Table 7 gives a brief overview of widely used measurement techniques and sensors that enable wearable technology.

Table 6: An overview of selected measurement techniques and sensor technologies.

Wearable device/sensor or Method	Measurement	Description
Photoplethysmography (PPG)	Heart rate, blood oxygen saturation levels (SpO2) [1], vascular resistance [2], blood pressure [22].	Uses one or two light sources (red or green) and a detector to capture volumetric changes associated with dilation and contraction of vessels in the dermis and hypodermis [3]. Worn on wrist and ear lobe.
Electrocardiogram (ECG)	Heart rate, heart rate variability, stroke volume	It is a voltage-time graph of the heart produced by electrodes which detect small changes in current due to systolic and diastolic heart function. Worn as a strap on the chest and limbs [4], as smart clothing [5].
Impedance Cardiogram (ICG)	Heart rate, cardiac output, stroke volume, ejection fraction	Electrodes placed at the neck and the diaphragm level, detect the changes of the thoracic impedance of blood and tissue caused by cardiac contraction. Accomplishes reproducible results of 97% [6].
Ballistocardiograph (BCG)	Ballistic reactions (e.g., forces, acceleration, displacement)	Measures ballistic reactions like displacement or acceleration resulting from the movement of blood due to expansion and contraction of the heart. There are different measurement systems with different interpretations: Starr BCG, Nickerson BCG, Dock BCG [7]. Sensors have the form of static charge-sensitive beds [9], piezoelectric films on chairs and beds [10, 11], piezoelectric films worn as soles under the feet [12], ear worn accelerometers [13], wrist worn accelerometers and chest seismocardiogram (SCG) [14]. New methods aim to understand BCG signals by using Computational Fluid Dynamics (CFD) and CT images from aortas [8].

Spirometry	Volume and velocity of air in each respiration cycle	It traditionally consists of an ergospirometer mask, and the use of ultrasonic transducers to measure the pressure difference in the mask. Recent efforts try to eliminate the use of a mask, by replacing it with miniaturized sensors near the mouth or nose in the form of humidity sensors [15], optical fiber sensors [16], MEMS capacitive pressure sensors [17], and textile sensors [18].
Respiratory Inductance Plethysmography (RIP)	Airway obstruction	Assesses pulmonary function based on the expansion and contraction of the chest and abdomen by detecting changes in the magnetic field induced by the worn wire loop [4]. Variations of RIP include elastomeric plethysmography [19] and impedance plethysmography [20].
Blood pressure oscillometer	Blood pressure	It measures pressure fluctuations in the cuff enabled by arterial's elasticity dependency on pressure [21]. It consists of multiple components: air-bladder cuff, pump, valves, pressure sensors, power supply etc.
Field effect transistors (FETs)	Body temperature	Temperature changes are translated into changes in the flow of current. Different FET sensors: graphene-based sensing elements, metal-polymer hybrids, and inorganic polymer hybrids [23- 26]. Attached to clothing, or skin mounted on wrist, ears, fingers etc.
Resistometry	Body Temperature	It relies on thermal expansion and changes of current flow to detect temperature change, by using different metal oxides [27, 28]. Skin mounted.
Accelerometer-based devices	Acceleration, reaction forces	They use piezoresistive, piezoelectric, or capacitive sensing elements to convert motion into an electrical signal [29]. Worn on different body parts or attached on clothing. There has been no standard way to present accurate physical activity and energy expenditure from the raw data [30, 31].
Enzymatic amperometric sensors	Different sweat biomarkers, like electrolytes (e.g., sodium and potassium ions) and metabolites (e.g., lactate [32] and glucose [33])	They use enzyme recognition due the formation of an oxidation-reduction reaction between the enzyme and the electrodes [34, 35]. Skin mounted.
Ion-selective electrode sensors (ISEs)	Different sweat biomarkers	They convert ion concentration to a voltage signal through direct potentiometry [35]. Different systems: solid-state membranes, liquid membranes, membranes on electrodes [36].

4.2 Data Mining for Wearable Sensors in Health Monitoring Systems

4.2.1 Data Mining Approach

This section may be reminiscent of some parts of section 2.2 since the data mining process of sensor data involves several steps and methods which are like image processing steps and methods. However, note that due to the temporal nature of the sensor data and the spatial nature of the image data, there are subtle differences between the methods applied. The main steps of any data mining approach involve 1) preprocessing of the raw data, 2) feature extraction and selection of useful information, and 3) applying learning models which get as input these features to perform tasks like anomaly detection, prediction and decision making [37], as shown in Fig. 7. The input data may be separated for training and testing the applied model before applying it to real world problems. Other parameters such as metadata, expert knowledge etc. may also improve the processes of feature extraction and training of the model [38].

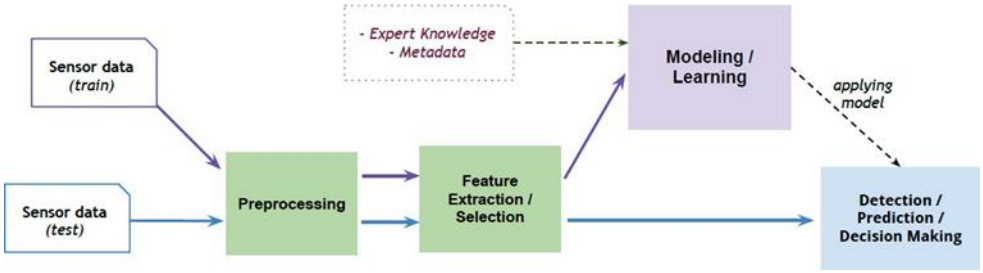

Figure 7. A generic architecture of the main data mining approach for wearable sensor data [37].

1. Preprocessing: This stage is necessary to filter out noise, artifacts, and other sensing errors from the raw data [39]. Different methods to filter sensor data from artifacts include threshold-based methods [40] and statistical tools for the interpolation of the missing data points [41]. To remove noise, methods in frequency domain are usually employed such as power spectral density (PSD) [42], fast Fourier transforms (FFT) [43], and low-pass/high-pass filtering tools. Gathering data from multiple sources requires also further treatment for issues such as formatting, normalization, and synchronization [39].

2. Feature Extraction and Selection: Feature extraction aims at extraction only the part of the input data that is useful to us. Signals can be analyzed in the time domain and in the frequency domain. Feature extraction in the time domain includes different statistical parameters attributed to the visible characteristics in data stream such as pick counts, mean, standard deviation etc. [44]. Analysis in the frequency domain is applied to extract information about the periodic behavior of time series in the data, and algorithms include power spectral density (PSD) [42], spectral density [45], wavelet coefficients of the signal [46], and low-pass/high-pass filters.

Feature selection aims to select a reduced set (or dimension) of the input data, which is representative of the original set. Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Independent Component Analysis (ICA) [47] are widely used methods for data dimension reduction, while other commonly used techniques in the literature include analysis of variance (ANOVA) [48], threshold-based rules [41], and Fourier transforms [43].

3. Classification: Input sensor data are classified to a relatively small number of general classes to make sense of them. Traditionally, statistical tools (e.g., mean, variance etc.) or statistical functions (e.g., risk function [41], factor analysis [49] etc.) are usually applied to simplify the data features and make formulations. Logistic regression (LR) [50] and multiple linear regression (MLR) [51] are two common statistical models. However, the continuous monitoring of sensors produces extremely large datasets that cannot be analyzed or interpreted using traditional data processing techniques. Therefore, to counteract such problem, machine-learning algorithms have been evolved to classify and interpret the results presenting the current state of the art in sensor informatics. Table 8 summarizes popular classification techniques of streaming sensor data.

Table 7: Selected learning methods for classification.

Method	Description	Benefits (O) / Limitations (•)
Support Vector Machines (SVM)	It is a binary classifier which constructs a high dimensional hyper-plane using quadratic programming for the separation of data points into two classes [52]. Using a kernel function, it can handle high dimensional data with a minimal training dataset. It also enables the build-in of expert knowledge by manipulating the kernel. It is commonly used for ECG, HR and SpO_2 signals.	○ Suitable for anomaly detection and decision-making tasks. • It cannot find unexpected information from unlabeled data. • Not appropriate to integrate metadata and other symbolic knowledge to enrich the results.
Decision Trees (DT)	The most robust features of input data are broken down into smaller subsets, represented in the form of tree nodes. Popular DT algorithms are ID3 [53], C4.5 [54] and J48 [55]. C4.5 estimates errors in the initial nodes and prunes the tree accordingly to make it more efficient, while J48 extends ID3's capabilities, by accounting for missing values, pruning noisy data, and deriving rules.	○ Simple and easy to implement. ○ They can handle data from multivariate sensors, with short stream of data. • They cannot find information which is hidden from the constructed features. • Not efficient when the number of input features gets large. • Prone to overfitting to the training dataset.
Random Forests (RF)	It consists of many different decisions trees by using bagging and feature randomness in a way that they are uncorrelated. The overall prediction of the trees is more accurate than any individual tree [56].	○ They outperform decision trees, though they show less accuracy than gradient boosted trees. ○ In contrast to decision trees, they are not prone to overfitting. ○ They do not require normalization. ○ Suitable for large datasets. • Overfitting risk. • Not appropriate for regression. • Biased towards variables with different level of attributes.
Gaussian Mixture Models (GMM)	They assume that input data is a linear combination of Gaussian distributions. After the initialization of some parameters, the model is re-estimated based on the input data and then makes statistical inferences about the properties of the sub-populations of the overall data [57].	○ It can detect unseen information in input data. ○ It is the fastest algorithm for learning mixture models. • Many points per mixture complicate the estimation of the covariance matrices, and unless one regularizes the

		covariances artificially, GMM will diverge and find solutions with infinite likelihood.
Hidden Markov Models (HMM)	Physiological data are modeled as Markov chain to compute the probability of each state's occurrence by calculating a histogram of the probabilities of successive states [58].	○ Convenient for modelling sequential data and detecting anomalies. ○ Hidden states can be inferred from other observations in the stream of data [59]. • Due to their Markovian nature, they do not consider the sequence of states leading into any given state. • They model the behavior of data using static distributions; hence they fail to model data which vary continuously with time.
Naive Bayes	It uses Bayes Theorem from statistics by assuming that the input features are conditionally independent [60].	○ The training time is linear with the training dataset. ○ Robust model regarding noise and missing values.
Bayesian Networks (BN)	It is a probabilistic graphical model comprising nodes and edges, which captures both conditionally dependent and conditionally independent relationships between random variables.	○ Ideal for taking an incident that occurred and predicting the likelihood that any of several potential causes was the tributary issue. • Its simple sophistication neglects data independency. • Large training datasets will not improve its prediction accuracy. • Not suitable for regression.
k-Nearest Neighbor (kNN)	It classifies unknown features based on similarity measures, e.g. distance function [61].	○ It requires no training. ○ Simple and easy to implement for multi-class problems. • Inefficient for huge datasets. • Unable to deal with missing values. • Outlier sensitive.
Neural Networks (NNs)	An artificial intelligence approach which mimics biological neural networks [62]. It is trained, i.e. it adjusts its network weights in each iteration of learning, based on the known classification of the training dataset. Common NNs are the multi-layer perceptron (MLP) [63], the deep MLP (DMLP) [64], the replicator neural network (RNN) [65], and long short-term memory units (LSTMs) [66].	○ Ideal for modelling non-linear systems [67]. ○ It can still improve on large training datasets. • The training process is time-consuming since NNs need to train on large datasets. • High computational cost. • Prone to overfitting.
Probabilistic Neural Networks (PNN)	It is based on Bayes theory and implements a statistical algorithm called kernel discriminant analysis. It differs from the back-propagation neural networks in classification problems without the need for massive forward and backward calculations [68].	○ Their training process is faster than conventional NNs. ○ They show no local minimum issues. ○ They can work with smaller datasets for training.

		• Slow execution of the network, and heavy memory requirements as the training data get larger.
Deep Neural Networks (DNNs)	It relies on the same theoretical foundations as NNs. However, deep learning accounts for the use of many hidden neurons and layers. The large number of neurons allows for the extensive coverage of the input data, while the layer-by-layer pipeline of non-linear combination of their outputs generates a lower dimensional projection of the input space. Convolutional Neural Nets (CNNs), Deep Belief Networks (DBNs) and stacked Autoencoders functioning as deep Autoencoders are among, if not, the most popular DL methods.	○ Features are automatically deduced and optimally tuned for desired outcome. ○ It offers robust results and the highest performance among any other method when applied to huge datasets. • With many layers of neurons, it gets so complex that one is unable to comprehend the output of the given input. • There is no standard theory in selecting the appropriate DDN method.

Below table 9 presents selected studies reported in the literature, where the above-mentioned classifications methods have been applied.

Table 8. Applications of modeling methods in monitoring with wearable sensors.

Year - [REF] Author	Parameters	Sensors	Methods	Description
2008 - [69] Hu et al.	Heart Rate (HR)	ECG	SVM	A binary classifier version of SVM was used to categorize ECG signals into normal and arrhythmia classes.
2010 - [70] F.T. Sun et al.	Motion, HRV, sweat electric resistance	ECG, GCR, 3D accelerometers	Three different methods: SVM, DT, BN	A continuous monitoring stress system based on physiological signals is presented. A comparative study is done for the three different classification algorithms. The best classification accuracy for 10-fold cross validation (92.4%) and between-subjects classification (80.9%) is obtained from using the DT and SVM classifier, respectively, with the all-feature combination.
2011 - [71] Fotiadis et al.	Mean HR, RR, SpO_2, Inhalation/Exhalation Duration, Body Temperature, etc.	-	SVM	In order to detect the patient's condition based on different vital signs, One-Against-All SVM approach with three different kernels (RBF, polynomial, and sigmoid) is employed in order to handle multi-label classification according to four different levels of severity.

2011 - [72] Thakker B., Vyas, A.L	Radial pulses	-	SVM	This work could identify gastritis and arthritis in a person using binary classifications of normal and abnormal radial pulses of ECG with the SVM algorithm. Frequency domain features derived from power spectral density of the pulse signal are ranked to achieve dimensionality reduction. Among the different kernels employed, the SVM with a linear kernel classifies the abnormal pulse signals with highest success rate of 9.2%.
2011 - [73] Wang et al.	HRV, inter-pulse interval (IPI)	ECG	GMM	A GMM method uses IPI signals of ECG to make secure the body sensor communications. The proposed system utilizes ECG signal behavior (which is unique for each person) as a signature for authenticating other knowledge (e.g., medication delivery content information).
2011 - [59] M. Tomizuka J. Bae	Gait motion	-	HMM	HMM was applied to analyze gait phases. For the detection of gait phases, the posterior probabilities from the HMM were utilized, and the transition matrix was analyzed to check the abnormal state transition between gait phases.
2011 - [74] Zhu, Ying	Blood Glucose	-	HMM	HMM is used to detect anomalies in blood glucose levels being measured. The learning of the HMM is done using historic data of normal measurements. The simulation results show that the applied technique is accurate in detecting anomalies in glucose levels and is robust (i.e., no false positives) in the presence of reasonable changes in the patient's daily routine.
2012 - [75] K.H. Lee et al.	Heart Rate	ECG	SVM	A SVM method was developed for detecting the arrhythmia and seizure episodes with ECG signals. This research showed that the formulation for the kernel function of the SVM method with polynomial transformations reduces substantially the real-time computations involved when compared to other kernels.
2012 - [76] Clifford, Q. Li	Blood flow pulses	PPG	MLP	An MLP network is applied to combine several individual signal quality metrics and physiological context and estimate the quality of the pulses in PPG. After putting several individual signal quality metrics as input the network optimizes the number of nodes (2–20) hidden layer in validation iterations.

				An accuracy of 97.5% on the training set and 95.2% on the test set was found.
2012 - [77] Clifton et al.	RR, HR, BP, SpO2	PPG, ECG	GMM	A patient-personalized system for analysis and inference is proposed. This framework used Gaussian process to estimate reliably the distribution of the values of the physiological data. The method has been developed to improve removing artifacts and missing data from individual subjects. The method is demonstrated using a large-scale clinical study in which 200 patients have been monitored using the proposed system.
2012 - [78] C. Bellos et al.	Motion, SpO2, HR, RR, temperature, etc.	ECG, PPG, respiration bands, 3D accelerometers, humidity and temperature sensors, microphone with context audio sensor	RF	A decision support system is developed to classify the severity of health level based on a multiple parameter set using RF classification as a version of the decision tree. For the construction of each tree of the forest, a new subset of the features was picked. For selecting the best tree, the method used threshold-based rules. The accuracy of the system has been checked with some predefined targets.
2013 - [79] Chatterjee et al.	Blood Glucose (BG)	-	RNN	In this work, the RNN is designed with 11 input variables, one output node as predicted BG level and three hidden layers each with 8 neurons. This network can predict blood glucose levels for the next day from accumulated data with an accuracy of 94%.
2013 - [47] Giri et al.	Heart Rate	ECG	Four different methods: SVM, GMM, PNN, kNN	The heart rate signals are decomposed into frequency sub-bands with a DWT, and a different algorithm: PCA, LDA or ICA is applied on the set of DWT coefficients to reduce the data dimension. Then, the selected features are fed into a different classification method: SVM, GMM, PNN or kNN. The results show that the ICA coupled with GMM classifier combination resulted in highest accuracy of 96.8%, sensitivity of 100% and specificity of 93.7% compared to other data reduction techniques (PCA and LDA) and classifiers.

2013 - [80] E. Gaura et al.	Posture, pulse, HR, multi-point skin temperatures, core temperature, CO_2	-	Two different methods: BN, DT	This study uses a multiple parameter set to predict heat stress. The two algorithms (BN and DT) used are trained on empirical data and have accuracies of 92.1 ± 2.9 and 94.4 ± 2.1%, respectively, when tested using leave-one-subject-out cross-validation.
2017 - [81] H Yin et al.	HR, BT, SpO2, RR, BP, BG etc.	ECG, GSR, etc.	BN, NB, kNN, J48, SVM, MLP, RF, etc.	A hierarchical health decision support system is proposed for disease diagnosis that integrates health data from wearable medical sensors (WMSs) into Computer-based clinical decision support systems CDSSs. The system offers impressive diagnostic accuracies for various diseases: arrhythmia (86 %), type-2 diabetes (78%), urinary bladder disorder (99%), renal pelvis nephritis (94%), and hypothyroid (95%). The authors estimate that the disease diagnosis modules of all known 69,000 human diseases would require just 62 GB of storage space in the WMS tier, making it capable for any present cloud station.
2017 - [82] Akmandor et al.	RR, BP, pulse etc.	ECG, pulse oximeter, GSR, etc.	SVM, kNN	An automatic stress detection and alleviation system, called SoDA, is presented. In the stress detection stage, SoDA achieves 95.8% accuracy with a distinct combination of supervised feature selection and unsupervised dimensionality reduction.
2018 - [83] E Katoch	HR, pulse, body temperature, motion, etc.	Accelerometers, temperature sensor, optical HR sensor, pulse sensor	Six different methods: SVM, BN, kNN, DT, LDA, NN	A method was developed for the automatic recognition of sedentary behavior related cardiovascular risk. NN topology: one neuron placed on the output layer, 12 neurons on the hidden layer and two neurons in the input layer. A comparative study using 10-fold cross-validation between the applied classification algorithms showed that SVM, NN, BN performed best with an accuracy of 95.00% ± 2.11%.
2018 - [84] MM Hassan et al.	Motion	Smartphone inertial sensors	Three different methods: SVM, DBN, NN	A robust human activity recognition (e.g., standing, walking etc.) system based on the smartphone sensors' data is proposed. DBN outperformed the other classification algorithms, achieving a mean recognition rate of 89.61% and an overall accuracy of 95.85%.

2018 - [85] F. Miao et al.	21 features from PPG and ECG signals	ECG, PPG	Three different methods: MLR, DT, NN	A framework for arterial stiffness monitoring is developed. Experimental results based on 501 diverse subjects showed that the MLR approach exhibited the best accuracy in vascular age estimation, while the pack propagation NN was best in cardiovascular disease risk estimation.
2021 - [86] ALI F. et al.	BG, BP, SpO2, HR, etc.	ECG, EEG	Eight different methods: Fuzzy classifier, CNN, LSTM, LR, kNN, SVM, MLP, RF	A framework is proposed that extracts different data types from multiple sources for patients with chronic diseases. A comparative study showed that LSTM achieved the highest accuracy among all the applied classification methods, (75%) for diabetes classification and (88%) in terms of BP classification.

4.2.2 Big Data Repositories, IoT, and Diagnostics

Towards the integration of data from multiple sensors, IoT technologies will facilitate the efficient exchange and gathering of information. According to [87], current IoT technologies include Radio-Frequency Identification (RFID) [88] used for device identification and tracking, Cloud Computing [89] for offering massive computer resources in terms of data storage and computing power, and nanotechnologies where sensors are in the scale of nanometers and their interconnection diminishes the utilization of a framework, called Internet of Nano-Things [90]. Currently, among these technologies, big data repositories may be deemed the most important since it allows machine learning algorithms to train with high efficiency on large datasets before making inferences for data. The pipeline of remote computing involves detection of pathogenic biomarkers in the human body from the biosensors (e.g., hypertension, diabetes etc.), and data processing with the help of machine learning in Cloud. Furthermore, developing analytical platforms can enable the analysis and linking of diverse datasets. An example is Apache Hadoop [91], which can allow for data to be spread across many severs with little reduction in performance. The back end in pipeline of remote computing may be a virtual assistant (VA), where human intervention is replaced by technology [92]. Sensely [93] is a software as a service-based device being used for regular checkup of patients with chronic disorders. It includes biosensors, machine-learning and telemedicine that connects the patients automatically to its clinicians upon noticing the threshold symptoms of disorder. The final aim of Big Data enterprise is to combine data from multiple sources, e.g. Electronic Health Records, imaging phenotypes, genomic data, etc., for predictive analytics and precision health medicine.

A novel healthcare monitoring framework is proposed in [86], which extracts large amounts of healthcare data from multiple sources (smartphones, wearable sensors, medical records, and social networks) to monitor efficiently patients with chronic conditions, warn patients before their health risk reaches a high level, and support physicians in offering better treatment plans. The framework comprises five different layers, namely the data collection layer, the data source layer, the data storage layer, the analytics engine layer, and the data representation layer. The data collection layer gathers data from various domains like wearable devices, doctor to patient discussions on social networks etc., while the source layer deals with data heterogeneity. The data storage layer is responsible for offloading to the Cloud server all the collected data through a wireless communications network. The analytics engine layer is

divided into two sub-layers: the data computation layer and the data classification layer. The data computation layer has different sub-models for tasks like data preprocessing, data dimensionality reduction and feature extraction, and word embedding. Note that ontology-based semantic knowledge, along with soft computing approaches, are employed to process and analyze the data for the extraction of required information. The classification layer uses Bi-LSTM, a machine learning approach, and utilizes ontologies for the classification of diabetes, BP, mental health, and drug side effects. By analyzing the multidimensional data about patients, this layer gets insights for the decision-making process. The proposed system uses Hadoop MapReduce with Machine Learning to reduce large-sized data about patient treatments. The representation layer, which is the last layer, presents the analysis results to physician who in turn suggests the appropriate treatment plan.

4.3 Applications of Wearable Sensor Technology in Healthcare

Wearable sensors enable the continuous monitoring of vital signs (e.g., temperature, heart rate etc.) and other secondary data such as motion, emotional state etc., promoting individuals to maintain a good state of health. Furthermore, by remote monitoring wearable sensors facilitate independent living in home environments or hospitals for patient management, and through anomaly detection and raising alarms they additionally ease managing a disease [94].

4.3.1 Maintenance of Health

4.3.1.1 Fall Identification and Prevention

Wearable sensors can serve the ever-growing elder population for the early detection and prevention of falls. Studies presenting wearable sensors for the early detection of falls have shown satisfactory results in laboratory settings, although their applicability may be limited in real-world situations [95]. E.g. in [96], a method detecting a fall at different phases with the help of tri-axial accelerometers, achieved 86.5%, 87.3% and 91.2% accuracy for fall detection at pre-impacts, impacts, and post-impacts, respectively. In [97], a novel hierarchical fall detection system using accelerometer sensors on the waist reported an accuracy of 99% in identifying falls. Lastly in [98], compressive sensing techniques were used to detect signals and binary tree classifies for evaluation, achieving 99% precision in identifying falls.

4.3.1.2 Physical Activity

Modern wearables can assist with behavior change interventions aiming to encourage individual physical activity, and thus having high impact on the society. For instance, by decreasing physical inactivity by 10%, more than 533,000 deaths could be averted every year [99]. A further study reported, reported a positive impact of wearable device-based system along with vibration reminders at 20-minute intervals that could change student posture during prolonged sedentary behavior [100]. However, when tracking exercise intensity wearable devices and algorithms are currently unable to track calory consumption, as demonstrated by Dooley et al. [101]. Lastly, it is common practice nowadays to assess a person's daily physical activity with wearable sensors and link it to their disease profile, e.g. in chronic kidney [102] and cardiovascular disease [103].

4.3.1.3 Early-Stage Cancer Identification

Recently, a European Consortium developed a novel miniaturized, portable device called SNIFFPHONE [104], named after the name of the consortium. This device enables the diagnosis of gastric cancer from exhaled breath, catching the disease at the very early stages where the survival rate is very high. The SNIFFPHONE device operates in two steps; first the ambient air is measured for reference and then the user exhales at a short distance from the inlet of the device. When the user exhales, the beginning and end of his/her breath is detected by a breath detector sensor. Then, the microfluidic system directs the breath sample to the sensors chamber, where the gas sensor array is located. The sensor array contains a chip with eight Gold Nanoparticles (GNP) gas sensors for the detection of volatile organic compound (VOC) biomarkers of gastrointestinal diseases with risk of developing cancer. The SNIFFPHONE sensors response resulting from the breath measurements along with other relevant user information is transferred wirelessly via the phone's internet to the cloud server for a remote analysis of the collected signals. Machine learning methods are then applied on the received data, while considering other clinical information of the same patient. Upon completion of the analysis, a clinical report including the diagnosis results is sent back to clinical doctor for diagnosis and tracking along with a brief feedback to the user. SNIFFPHONE analysis perform with an 93.4% accuracy.

4.3.2 Patient Management

4.3.2.1 Patients with Cancer

Advances in cancer therapeutics have improved the survival rate and quality of life in patients affected by various cancers, but have been accompanied by treatment related cardiotoxicity, e.g. left ventricular (LV) dysfunction and/or overt heart failure (HF). In their work [105], the authors have demonstrated the potential of using wearable seismocardiography (SCG) to assess the clinical status of patients with LV relaxation dysfunction, when monitoring cancer treatment-related cardiovascular toxicity in patients undergoing cancer treatment. In another study [106], thirty-seven patients were evaluated (54% male, median 62 years) to ascertain associations between wearable activity monitor metrics (steps, distance, stairs) and performance status, clinical outcomes (adverse events, hospitalization times, survival rates), and patient-reported outcomes (PROs) using correlation statistics and multivariable logistic regression models. Patients averaged 3700 steps, 1.7 miles, and 3 flights of stairs per day. Each 1000 steps/day increase was associated with reduced odds for adverse events, hospitalizations, and hazard for death. Significant correlations were also observed between activity metrics and PROs.

4.3.2.2 Patients with Stroke

Stroke is a major cause of acquired disability in the global population. In stroke rehabilitation, digital biomarkers (e.g. activity levels or postures) could provide clinicians and patients interpretable feedback besides estimated clinical assessment scores and help devising personalized therapy recommendations based on continuous measurement. In their work [107], Adrian Derungs et al. show that wearable sensors and digital biomarkers offer opportunities to investigate changes during the recovery process in patients after stroke and they propose three novel digital biomarkers for longitudinal, bilateral movement evaluation, which are viable for therapy and free-living. In another study by Burridge et al. [108], the researchers developed a wearable device with embedded inertial and mechanomyography sensors, algorithms to classify functional movement, and a graphical user interface to present meaningful data to patients to support a home exercise program.

4.3.2.3 Chronic Pulmonary Patients

Chronic obstructive pulmonary disease is a type of lung disease caused by poor airflow that makes breathing difficult. As a chronic malady, it typically worsens over time, so extensive, long-term pulmonary rehabilitation exercises and patient management are required. A team of researchers designed a system which provides a comfortable and cost-effective option for the remote rehabilitation of patients with chronic breathing difficulties [109]. In this system, data regarding motion captured by a stereo-camera are fused with the signals from a PPG sensor and they are fed as input variables to an evaluation framework. In addition, the system included a set of rehabilitation exercises specific for pulmonary patients, and provided exercise tracking progress, patient performance, exercise assignments, and exercise guidance.

4.3.3 Disease Management

4.3.3.1 Heart Disorders

Implantable Cardioverter Defibrillator (ICD) is a battery-powered device that keeps track of the heart rate for people with heart disorders. In addition, if an abnormal heart rhythm (e.g. arrhythmia) is detected the device will deliver an electric shock to restore a normal heartbeat with the help of wires connecting the ICD to the heart. ICD can reduce sudden arrhythmic death in patients who are at high risk [110]. An alternate option in case of temporary inability to implant an ICD, and lastly refusal of an indicated ICD by the patient, is the wearable cardioverter defibrillator (WCD) indicated to prevent sudden arrhythmic death. Primarily, the WCD is designed to detect and treat automatically ventricular tachyarrhythmias. An WCD consists of tantalum oxide electrodes for long-term electrocardiogram (ECG) monitoring and has the characteristics of an ICD, but does not need to be implanted, and it has similarities with an external defibrillator, but does not require a bystander to apply lifesaving shocks when necessary [111].

Further studies, see e.g. [112], investigated the efficacy of a wireless digital watch for remote monitoring for the vital signs of patients, compared to traditional clinical monitors. The overall agreement between the watch and clinical monitors was statistically significant, and the wearable device provided reliable heart rate data for about 80% of the patients. In a similar study [113], the researchers showed that a wrist-worn personal fitness tracker device could be used to monitor the heart rate of patients, although with a systematic error, that being, the heart rate was slightly lower than the conventional method of continuous electrocardiographic (cECG) monitoring.

4.3.3.2 Blood Disorders

Wearable devices can improve hypertension control and medication adherence through easier logging of repeated blood pressure measurements, better connectivity with healthcare provides, and medication reminder alerts [114]. Current tools for the out-the-office measurements include wireless upper arm blood pressure cuffs and cuffless devices. Wireless upper arm cuffs are automated oscillometric devices that synchronize via bluetooth technology to a computer or smartphone. These work the same way as conventional clinical devices by recording vibrations in the arterial wall to establish systolic and diastolic pressure. Blood pressures are automatically logged and saved. The cuffless devices are applied to the wrist or finger, measuring via optical sensors beat-to-beat variability, and via mathematic formulations they can compute systolic and diastolic readings. For instance, one such device, Somnotouch-NIBP, uses finger PPG and three ECG leads connected to a watch-like control unit to obtain systolic and diastolic blood pressure via pulse wave velocity measurements [115]. The benefits of these blood pressure sensors

include the ability to monitor continuously and avoid sleep-disrupting cuff inflations when measurements are required at night. However, the tradeoff for this convenience lies in the accuracy of the devices. Although many wireless sensors have been validated and FDA approved for clinical use, measurements can vary as much as 20 mmHg from blood pressures derived using brachial cuff [116]. While there is inadequate evidence to recommend cuffless devices to patients at present, there are over 1000 clinical trials currently registered with www.clinicaltrials.gov to evaluate the feasibility, accuracy, and safety of various sensor technologies. Other devices for pressure blood measurements have been tried on patients with orthostatic hypotension, who have pathologic hemodynamics related to changes in body posture. Researchers designed a new cephalic laser blood flowmeter that can be worn on the tragus to investigate hemodynamics upon rising from a sitting or squatting posture. This new wearable cerebral blood flow (CBF) meter is potentially useful for estimating cephalic hemodynamics and objectively diagnosing cerebral ischemic symptoms of patients in a standing posture [117].

4.3.3.3 Diabetes

People with diabetes have a deficiency of the insulin production by the pancreas, which prevents a correct metabolism of glucose. Current methods of managing diabetes include insulin injections by some external device. Cutting the supply or incorrect doses of insulin, may initiate a chain of subsequent reactions, possibly even leading to life-threatening health conditions. The amount of insulin needed varies, depending mostly on the nutrition and activity of the patient. People with diabetes must therefore continuously measure their glycemia and perform several subcutaneous injections of insulin per day. Through subcutaneous insulin infusion, mathematical models and computer simulation of the human metabolic system, real-time continuous glucose monitoring (CGM), and control algorithms driving closed-loop control systems known as the “artificial pancreas” [118], the quality of life of many people suffering from diabetes type I can be significantly enhanced. A common technique utilized by most of the commercialized CGM systems is the glucose-oxidase electrochemical principle [119], which makes use of a minimally invasive needle sensor, usually inserted in the subcutaneous tissue, in the abdomen or on the arm, to measure an electrical current signal generated by the glucose-oxidase reaction.

However, in recent years, new commercialized state-of-the-art sensors start to emerge and alter the scope of glucose measuring invasive techniques. Two current consumer technologies include: Glutrac and AerBetic. Glutrac [120] is a smartwatch that claims to provide non-invasive continuous glucose monitoring with the use of optical sensors, while still achieving a high level of accuracy. This smartwatch records health data every 15 minutes. Then it analyzes this data in the cloud and with the help of AI algorithms it provides the precise blood glucose prediction of the current measurement [121]. AerBetic [122] is also a smartwatch, designed based on the idea that dogs can smell a person’s blood sugar fluctuations. To monitor the changes in blood sugar levels, AerBetic uses a nano-sensor that detects gases humans emit. This sensor is sensitive enough that it detects gases at parts per billion level, so the individual doesn’t have to directly breath to the watch.

Other novel technologies involve glucose measurements via sweat. In a reported study [123], researchers explore how skin patches can use the glucose levels in sweat as an indicator of overall blood glucose concentration and deliver insulin in a non-invasive manner. Their device captures sweat from the person’s skin, and sensors within the patch measure sweat’s pH level and temperature changes. Once high sugar levels are recognized, built-in heaters in the patch dissolve a layer of coating, exposing microneedles that release a drug (metformin) that can regulate and reduce high blood sugar levels. Blood sugar readings are also wirelessly transmitted to a mobile device so that long term trends are simple to read and monitor.

4.3.3.4 Parkinson's Disease

Parkinson's disease (PD) is the second-most common neurodegenerative disease and a major cause of disability worldwide [124]. Currently, treatment is based on subjective questionnaires and rare patient doctor interactions. Wearable devices are competent to collect useful data that offer insights into the diagnosis and the effects of treatment interventions to manage Parkinson's disease. E.g., bradykinesia is a primary symptom of PD. In a study [125], researchers developed a wearable device to assess the severity of the Parkinsonian bradykinesia based on the ten-second whole-hand-grasp action, and a regression model to fit the parameters under consideration. In their work the authors claim that the proposed quantification model demonstrated greater goodness-of-fit when compared with related works. With dyskinesia being another characteristic symptom of PD, researchers developed an objective dyskinesia score by using a motion capture system to collect patient kinematic data [126]. Furthermore, Freezing of Gait (FoG) is a common symptom in PD occurring with significant variability, and severity and is associated with increased risk of falls. Recently, Fotiadis et al. [127] validated a novel system based on a pair of pressure insoles equipped with a 3D accelerometer to detect FoG episodes. In their study, twenty PD patients attended a motor assessment protocol organized into eight multiple videos recorded sessions, both in clinical and ecological settings and both in the ON and OFF state. Then the researchers compared the FoG episodes detected using the processed data gathered from the insoles with those tagged by a clinician on video recordings. Their algorithm correctly detected 90% of the episodes.

Finally, a fully integrated commercialized technology for PD monitoring is the PDMonitor®. PD Monitor consists of: 1) five monitoring devices worn by the patient on different body parts, where each device collects 3D kinematic data, 2) a docking station (SmartBox) for charging of monitoring devices and uploading patient information to the cloud, 3) a mobile app for patients or caregivers to interact with the device and provide important diary information, 4) the Cloud, where patient data is securely stored and 5) the Physician Tool, a web-based application to view and download patient reports with a comprehensive and objective assessment of the PD symptoms. PDMonitor can monitor bradykinesia, dyskinesia, tremor, FoG, gait disturbances and postural instability, to name a few.

4.3.3.5 Emotional Health State

It is well recognized that emotions impact the overall health state of an individual. Emotions are neural responses to internal and/or external events that may manifest negatively as depression, anxiety, stress, fatigue, sleepiness, sleep disorders etc. Different chemical biomarkers indicate the emotional state of a person, and thus wearable devices have been proposed, which through chemical analysis (e.g., cortisol, prolactin, hGH, ACTH, and lactate) of saliva and blood can measure, e.g. stress state [128, 129]. Emotional state can also be inferred from measurement of various physiological parameters including heart rate [130], heart rate variability [131], respiration rate [132], blood pressure [133], electroencephalogram (EEG) [134], electromyogram (EMG) and electro-oculogram (EOG) [135], plethysmograph (PPG) [136], galvanic skin response (GSR) [137], and skin temperature [130]. In addition, night sleep patterns can be a prominent indicator of stress. In a study [138], the results showed that stress disrupts night sleep, and thus causing sympathetic predominance which then can be used to quantify stress levels. Finally, the advent of smartphones offers a range of emotion detection capabilities, since they can utilize an ever-increasing number of apps that detect and respond to end user emotional states [139, 140].

4.4 Future Landscape

4.4.1 Selective Fusion of Multiple Signals

Effective approaches are needed for the integration of different physiological signals from multiple sources [141]. E.g., in heartbeat monitoring the use of different physiological signals can improve robustness and accuracy of detection. In [142], the authors propose an approach based on CNNs to fuse various physiological signals and enhance heartbeat detection. In another study [143], a deep fusional attention neural network, named FusionAtt, could learn channel-aware representations of multi-channel biosignals. This system outperformed current state-of-the-art approaches in two clinical tasks: seizure detection using data from 23-channel electroencephalogram signals and sleep stage classification using data from 14-channel polysomnography.

4.4.2 Improvements in Model Training

Currently, several efforts are focused on the challenge of getting massive quantities of reliable and consistent labels that can be used to train machine learning algorithms [141]. In their study [144], researchers provide a systematic approach employing Bayesian methods for the automated labeling, and fusion of different physiological signals to support decision-making in personalized care. A study addressing the problem manual annotations in digital pathology [145], presented a multiple instance learning-based approach to train deep neural networks that generate semantically rich tile-level features. These features are then used as input into an RNN to integrate the information across the whole slide and report the final classification result. With the proposed system, pathologists could exclude 65–75% of slides while retaining 100% sensitivity.

4.4.3 Novel Approaches to Handle Longitudinal Data

Yildirim et al., [146] describe a way to compress signals from Holter monitor devices using a deep convolutional auto-encoder. The compressed electrocardiogram signals are then fed into LSTM classifiers to detect arrhythmia. Based on the compressed signals, storage requirement and classification time were reduced, while studies conducted with MIT-BIH arrhythmia database, reported classification accuracy of over 99%. In medical image analysis, image registration is a frequent task in medical imaging and computer-aided diagnosis.

4.4.4 Self-powered and battery-free wearable systems

Long-term stability in energy requirements will play a major role in wearable devices. Traditional batteries fail to meet the energy requirements of storage units in wearable devices; hence several energy harvesting solutions have been proposed to address the limitations of the bulky batteries. Among the proposed solutions, there have been systems that exploit solar energy, mechanical and thermal energy (produced by the subject). Regarding solar energy, a solar cell is developed on a plastic substrate to convert the incoming light into electricity, and supply power to the whole sensing system, Fig. 8a. Regarding mechanical energy, a magneto-mechanotriboelectric nanogenerator which generates electricity from the alternating magnetic field has been reported recently, successfully powering up an indoor wireless positioning system Fig. 8b. Regarding thermal energy, a flexible thermoelectric generator (TEG) is developed with a polymer-based heat sink assembled on the top surface to further increase the output power density from 8 to 38 μW cm^{-2}. An electrocardiography (ECG) sensing circuit is also fabricated on a flexible

PCB substrate and powered by the wearable TEG using body heat as the power source Fig. 8c. A systematic review of the aforementioned propositions, and other that relate to implanted devices can be found in [147].

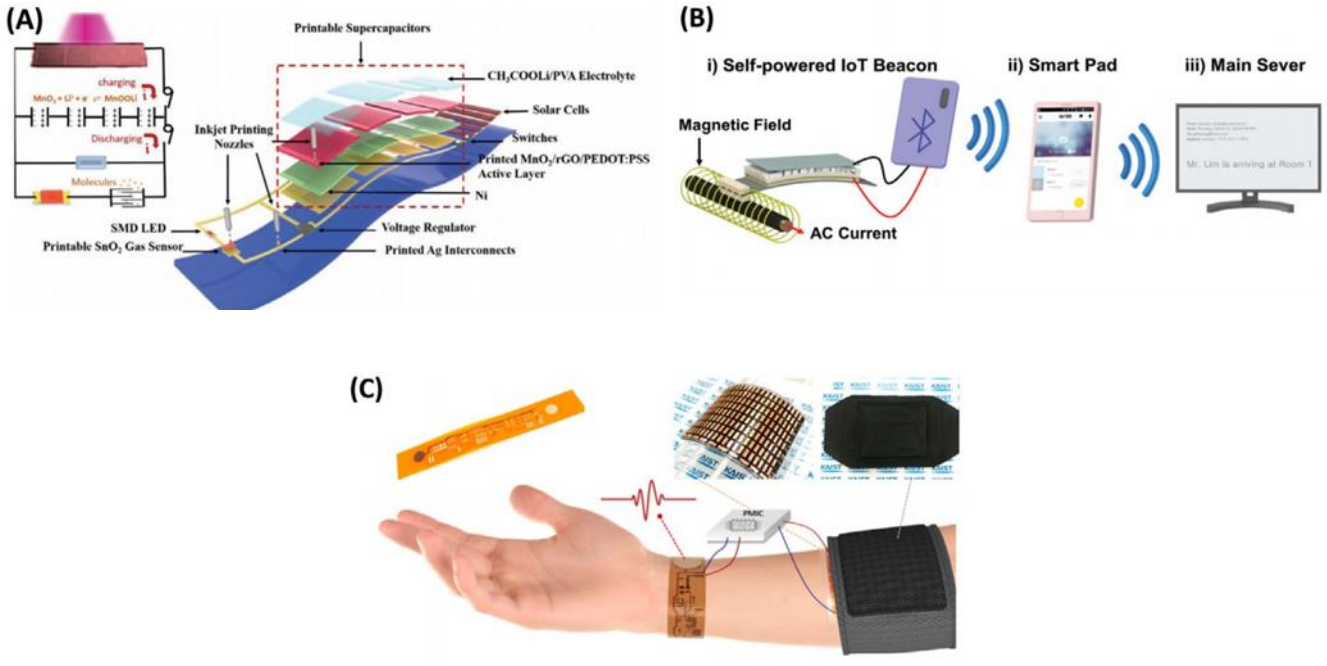

Figure 8: A. Self-powered gas monitoring system with embedded solar cells as the energy source; **B**. Self-powered indoor IoT positioning system integrated with energy harvesting and storage units; **C**. Self-powered wearable electrocardiography system powered by a wearable thermoelectric generator [147].

4.5 References

[1] D. J. Faber, M. C. G. Aalders, E. G. Mik, B. A. Hooper, M. J. C. van Gemert, and T. G. van Leeuwen, "Oxygen saturation-dependent absorption and scattering of blood," *Phys. Rev. Lett.*, vol. 93, no. 2, p. 28102, 2004.

[2] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, p. R1, 2007.

[3] H. H. Asada, P. Shaltis, A. Reisner, S. Rhee, and R. C. Hutchinson, "Mobile monitoring with wearable photoplethysmographic biosensors," *IEEE Eng. Med. Biol. Mag.*, vol. 22, no. 3, pp. 28–40, 2003.

[4] I. cheol Jeong, D. Bychkov, and P. C. Searson, "Wearable devices for precision medicine and health state monitoring," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1242–1258, 2018.

[5] Y.-D. Lee and W.-Y. Chung, "Wireless sensor network based wearable smart shirt for ubiquitous health and activity monitoring," *Sensors Actuators B Chem.*, vol. 140, no. 2, pp. 390–395, 2009.

[6] J. Fortin *et al.*, "Non-invasive beat-to-beat cardiac output monitoring by an improved method of

transthoracic bioimpedance measurement," *Comput. Biol. Med.*, vol. 36, no. 11, pp. 1185–1203, 2006.

[7] W. R. Scarborough *et al.*, "Proposals for ballistocardiographic nomenclature and conventions: Revised and extended: Report of committee on ballistocardiographic terminology," *Circulation*, vol. 14, no. 3, pp. 435–450, 1956.

[8] L. Giovangrandi, O. T. Inan, R. M. Wiard, M. Etemadi, and G. T. A. Kovacs, "Ballistocardiography—a method worth revisiting," in *2011 annual international conference of the IEEE engineering in medicine and biology society*, 2011, pp. 4279–4282.

[9] J. Alihanka, K. Vaahtoranta, and I. Saarikivi, "A new method for long-term monitoring of the ballistocardiogram, heart rate, and respiration," *Am. J. Physiol. Integr. Comp. Physiol.*, vol. 240, no. 5, pp. R384–R392, 1981.

[10] G. S. Chung, J. S. Lee, S. H. Hwang, Y. K. Lim, D.-U. Jeong, and K. S. Park, "Wakefulness estimation only using ballistocardiogram: Nonintrusive method for sleep monitoring," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 2459–2462.

[11] C. Bruser, K. Stadlthanner, S. de Waele, and S. Leonhardt, "Adaptive beat-to-beat heart rate estimation in ballistocardiograms," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 5, pp. 778–786, 2011.

[12] J. Zhao and Z. You, "A shoe-embedded piezoelectric energy harvester for wearable sensors," *Sensors*, vol. 14, no. 7, pp. 12497–12510, 2014.

[13] D. Da He, E. S. Winokur, and C. G. Sodini, "An ear-worn continuous ballistocardiogram (BCG) sensor for cardiovascular monitoring," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 5030–5033.

[14] M. Etemadi and O. T. Inan, "Wearable ballistocardiogram and seismocardiogram systems for health and performance," *J. Appl. Physiol.*, vol. 124, no. 2, pp. 452–461, 2018.

[15] U. Mogera, A. A. Sagade, S. J. George, and G. U. Kulkarni, "Ultrafast response humidity sensor using supramolecular nanofibre and its application in monitoring breath humidity and flow," *Sci. Rep.*, vol. 4, no. 1, pp. 1–9, 2014.

[16] Y. Kang, H. Ruan, Y. Wang, F. J. Arregui, I. R. Matias, and R. O. Claus, "Nanostructured optical fibre sensors for breathing airflow monitoring," *Meas. Sci. Technol.*, vol. 17, no. 5, p. 1207, 2006.

[17] M. Chattopadhyay and D. Chakraborty, "A new scheme for determination of respiration rate in human being using MEMS Based capacitive pressure sensor," in *Next Generation Sensors and Systems*, Springer, 2016, pp. 143–160.

[18] P. Rai, S. Oh, P. Shyamkumar, M. Ramasamy, R. E. Harbaugh, and V. K. Varadan, "Nano-bio-textile sensors with mobile wireless platform for wearable health monitoring of neurological and cardiovascular disorders," *J. Electrochem. Soc.*, vol. 161, no. 2, p. B3116, 2013.

[19] I. Levai, V. Sidoroff, and R. Iles, "An Introduction to the non-invasive non-contact assessment of respiratory function," *Respir. Ther.*, vol. 7, no. 5, p. 43, 2012.

[20] G. B. Moody, R. G. Mark, A. Zoccola, and S. Mantero, "Derivation of respiratory signals from multi-lead ECGs," *Comput. Cardiol.*, vol. 12, no. 1985, pp. 113–116, 1985.

[21] L. A. Geddes, M. Voelz, C. Combs, D. Reiner, and C. F. Babbs, "Characterization of the oscillometric method for measuring indirect blood pressure," *Ann. Biomed. Eng.*, vol. 10, no. 6, pp. 271–280,

1982.

[22] C. Z. Myint, K. H. Lim, K. I. Wong, A. A. Gopalai, and M. Z. Oo, "Blood pressure measurement from photo-plethysmography to pulse transit time," in *2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES)*, 2014, pp. 496–501.

[23] H. Jin, Q. Jin, and J. Jian, "Smart materials for wearable healthcare devices," in *Wearable Technologies*, IntechOpen, 2018.

[24] T. Q. Trung, S. Ramasundaram, B. Hwang, and N. Lee, "An all-elastomeric transparent and stretchable temperature sensor for body-attachable wearable electronics," *Adv. Mater.*, vol. 28, no. 3, pp. 502–509, 2016.

[25] X. Wang, L. Dong, H. Zhang, R. Yu, C. Pan, and Z. L. Wang, "Recent progress in electronic skin," *Adv. Sci.*, vol. 2, no. 10, p. 1500169, 2015.

[26] M. Kaltenbrunner *et al.*, "An ultra-lightweight design for imperceptible plastic electronics," *Nature*, vol. 499, no. 7459, pp. 458–463, 2013.

[27] Z. Popovic, P. Momenroodaki, and R. Scheeler, "Toward wearable wireless thermometers for internal body temperature measurements," *IEEE Commun. Mag.*, vol. 52, no. 10, pp. 118–125, 2014.

[28] M. Huang, T. Tamura, Z. Tang, W. Chen, and S. Kanaya, "A wearable thermometry for core body temperature measurement and its experimental verification," *IEEE J. Biomed. Heal. informatics*, vol. 21, no. 3, pp. 708–714, 2016.

[29] C.-C. Yang and Y.-L. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring," *Sensors*, vol. 10, no. 8, pp. 7772–7788, 2010.

[30] S. L. Murphy, "Review of physical activity measurement using accelerometers in older adults: considerations for research design and conduct," *Prev. Med. (Baltim).*, vol. 48, no. 2, pp. 108–114, 2009.

[31] D. S. Ward, K. R. Evenson, A. Vaughn, A. B. Rodgers, and R. P. Troiano, "Accelerometer use in physical activity: best practices and research recommendations.," *Med. Sci. Sports Exerc.*, vol. 37, no. 11 Suppl, pp. S582-8, 2005.

[32] S. Anastasova *et al.*, "A wearable multisensing patch for continuous sweat monitoring," *Biosens. Bioelectron.*, vol. 93, pp. 139–145, 2017.

[33] J. Kim, A. S. Campbell, and J. Wang, "Wearable non-invasive epidermal glucose sensors: A review," *Talanta*, vol. 177, pp. 163–170, 2018.

[34] S. Datta, L. R. Christena, and Y. R. S. Rajaram, "Enzyme immobilization: an overview on techniques and support materials," *3 Biotech*, vol. 3, no. 1, pp. 1–9, 2013.

[35] M. Chung, G. Fortunato, and N. Radacsi, "Wearable flexible sweat sensors for healthcare monitoring: a review," *J. R. Soc. Interface*, vol. 16, no. 159, p. 20190217, 2019.

[36] D. Ammann, *Ion-selective microelectrodes: principles, design and application*, vol. 50. Springer Science & Business Media, 2013.

[37] H. Banaee, M. U. Ahmed, and A. Loutfi, "Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges," *Sensors*, vol. 13, no. 12, pp. 17472–17500, 2013.

[38] J. Gialelis, P. Chondros, D. Karadimas, S. Dima, and D. Serpanos, "Identifying chronic disease complications utilizing state of the art data fusion methodologies and signal processing algorithms," in *International Conference on Wireless Mobile Communication and Healthcare*, 2011, pp. 256–263.

[39] D. Sow, D. S. Turaga, and M. Schmidt, "Mining of sensor data in healthcare: a survey," *Manag. Min. Sens. data*, pp. 459–504, 2013.

[40] Y. Mao, W. Chen, Y. Chen, C. Lu, M. Kollef, and T. Bailey, "An integrated data mining approach to real-time clinical monitoring and deterioration warning," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1140–1148.

[41] D. Apiletti, E. Baralis, G. Bruno, and T. Cerquitelli, "Real-time analysis of physiological data to support medical applications," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 3, pp. 313–321, 2009.

[42] A. Buchwald and K. W. Martin, "Mathematical Preliminaries: Power Spectral Densities of Random Data and Noise," in *Integrated Fiber-Optic Receivers*, Springer, 1995, pp. 27–104.

[43] P. P. G. Dyke, *An introduction to Laplace transforms and Fourier series*. Springer, 2014.

[44] R. R. Singh, S. Conjeti, and R. Banerjee, "An approach for real-time stress-trend detection using physiological signals in wearable computing systems for automotive drivers," in *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2011, pp. 1477–1482.

[45] J. Fan and Q. Yao, "Spectral density estimation and its applications," *Nonlinear Time Ser. Nonparametric Parametr. Methods*, pp. 275–312, 2003.

[46] B. Delyon and A. Juditsky, "Estimating wavelet coefficients," in *Wavelets and Statistics*, Springer, 1995, pp. 151–168.

[47] D. Giri *et al.*, "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform," *Knowledge-Based Syst.*, vol. 37, pp. 274–282, 2013.

[48] B. Shahbaba, "Analysis of Variance (ANOVA)," in *Biostatistics with R*, Springer, 2012, pp. 221–234.

[49] G. N. Pradhan, R. Chattopadhyay, and S. Panchanathan, "Processing body sensor data streams for continuous physiological monitoring," in *Proceedings of the international conference on Multimedia information retrieval*, 2010, pp. 479–486.

[50] B. F. French, J. C. Immekus, and H.-J. Yen, "Logistic regression," in *Handbook of quantitative methods for educational research*, Brill Sense, 2013, pp. 145–165.

[51] L. S. Aiken, S. G. West, S. C. Pitts, A. N. Baraldi, and I. C. Wurpts, "Multiple linear regression," *Handb. Psychol. Second Ed.*, vol. 2, 2012.

[52] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[53] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[54] J.-Y. Yeh, T.-H. Wu, and C.-W. Tsao, "Using data mining techniques to predict hospitalization of hemodialysis patients," *Decis. Support Syst.*, vol. 50, no. 2, pp. 439–448, 2011.

[55] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," *Proc. Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 6, 2013.

[56] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble machine learning*, Springer, 2012, pp. 157–175.

[57] D. A. Reynolds, "Gaussian Mixture Models.," *Encycl. biometrics*, vol. 741, pp. 659–663, 2009.

[58] O. Thomas *et al.*, "Wearable sensor activity analysis using semi-Markov models with a grammar," *Pervasive Mob. Comput.*, vol. 6, no. 3, pp. 342–350, 2010.

[59] J. Bae and M. Tomizuka, "Gait phase analysis based on a Hidden Markov Model," *Mechatronics*, vol. 21, no. 6, pp. 961–970, 2011.

[60] G. I. Webb, "Naïve Bayes.," *Encycl. Mach. Learn.*, vol. 15, pp. 713–714, 2010.

[61] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, "K-nearest neighbor classification," in *Data mining in agriculture*, Springer, 2009, pp. 83–106.

[62] M. Paliwal and U. A. Kumar, "Neural networks and statistical techniques: A review of applications," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 2–17, 2009.

[63] M. Aitkin and R. Foxall, "Statistical modelling of artificial neural networks using the multi-layer perceptron," *Stat. Comput.*, vol. 13, no. 3, pp. 227–239, 2003.

[64] N. Masih, H. Naz, and S. Ahuja, "Multilayer perceptron based deep neural network for early detection of coronary heart disease," *Health Technol. (Berl).*, pp. 1–12, 2020.

[65] H. A. Dau, V. Ciesielski, and A. Song, "Anomaly detection using replicator neural networks trained on examples of one class," in *Asia-Pacific Conference on Simulated Evolution and Learning*, 2014, pp. 311–322.

[66] G. Van Houdt, C. Mosquera, and G. Napoles, "A review on the long short-term memory model," *Artif. Intell. Rev.*, vol. 53, pp. 5929–5955, 2020.

[67] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis." Elsevier, 2013.

[68] B. Mohebali, A. Tahmassebi, A. Meyer-Baese, and A. H. Gandomi, "Probabilistic neural networks: a brief overview of theory, implementation, and application," *Handb. Probabilistic Model.*, pp. 347–367, 2020.

[69] F. Hu, M. Jiang, L. Celentano, and Y. Xiao, "Robust medical ad hoc sensor networks (MASN) with wavelet-based ECG data mining," *Ad Hoc Networks*, vol. 6, no. 7, pp. 986–1012, 2008.

[70] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-aware mental stress detection using physiological sensors," in *International conference on Mobile computing, applications, and services*, 2010, pp. 282–301.

[71] C. Bellos, A. Papadopoulos, R. Rosso, and D. I. Fotiadis, "A support vector machine approach for categorization of patients suffering from chronic diseases," in *International Conference on Wireless Mobile Communication and Healthcare*, 2011, pp. 264–267.

[72] B. Thakker and A. Lal Vyas, "Support vector machine for abnormal pulse classification," *Int. J. Comput. Appl.*, vol. 22, no. 7, pp. 13–19, 2011.

[73] W. Wang, H. Wang, M. Hempel, D. Peng, H. Sharif, and H.-H. Chen, "Secure stochastic ECG signals based on Gaussian mixture model for $ e $-healthcare systems," *IEEE Syst. J.*, vol. 5, no. 4, pp. 564–573, 2011.

[74] Y. Zhu, "Automatic detection of anomalies in blood glucose using a machine learning approach," *J. Commun. Networks*, vol. 13, no. 2, pp. 125–131, 2011.

[75] K. H. Lee, S.-Y. Kung, and N. Verma, "Low-energy formulations of support vector machine kernel

functions for biomedical sensor applications," *J. Signal Process. Syst.*, vol. 69, no. 3, pp. 339–349, 2012.

[76] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiol. Meas.*, vol. 33, no. 9, p. 1491, 2012.

[77] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, 2012.

[78] C. Bellos, A. Papadopoulos, R. Rosso, and D. I. Fotiadis, "Categorization of patients' health status in COPD disease using a wearable platform and random forests methodology," in *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*, 2012, pp. 404–407.

[79] S. Chatterjee, K. Dutta, H. Xie, J. Byun, A. Pottathil, and M. Moore, "Persuasive and pervasive sensing: A new frontier to monitor, track and assist older adults suffering from type-2 diabetes," in *2013 46th Hawaii international conference on system sciences*, 2013, pp. 2636–2645.

[80] E. Gaura, J. Kemp, and J. Brusey, "Leveraging knowledge from physiological data: On-body heat stress risk prediction with sensor networks," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 6, pp. 861–870, 2013.

[81] H. Yin and N. K. Jha, "A health decision support system for disease diagnosis based on wearable medical sensors and machine learning ensembles," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 3, no. 4, pp. 228–241, 2017.

[82] A. O. Akmandor and N. K. Jha, "Keep the stress away with SoDA: Stress detection and alleviation system," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 3, no. 4, pp. 269–282, 2017.

[83] E. Kańtoch, "Recognition of sedentary behavior by machine learning analysis of wearable sensors during activities of daily living for telemedical assessment of cardiovascular risk," *Sensors*, vol. 18, no. 10, p. 3219, 2018.

[84] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Futur. Gener. Comput. Syst.*, vol. 81, pp. 307–313, 2018.

[85] F. Miao, X. Wang, L. Yin, and Y. Li, "A wearable sensor for arterial stiffness monitoring based on machine learning algorithms," *IEEE Sens. J.*, vol. 19, no. 4, pp. 1426–1434, 2018.

[86] F. Ali *et al.*, "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Futur. Gener. Comput. Syst.*, vol. 114, pp. 23–43, 2021.

[87] A. Kumar, A. O. Salau, S. Gupta, and K. Paliwal, "Recent trends in IoT and its requisition with IoT built engineering: A review," *Adv. Signal Process. Commun.*, pp. 15–25, 2019.

[88] M. Goshey, "Radio Frequency Identification (RFID)." Springer, 2008.

[89] N. Antonopoulos and L. Gillam, *Cloud computing*. Springer, 2010.

[90] N. Akhtar and Y. Perwej, "The internet of nano things (IoNT) existing state and future Prospects," *GSC Adv. Res. Rev.*, vol. 5, no. 2, pp. 131–150, 2020.

[91] L. Belcastro, F. Marozzo, and D. Talia, "Programming models and systems for big data analysis," *Int. J. Parallel, Emergent Distrib. Syst.*, vol. 34, no. 6, pp. 632–652, 2019.

[92] M. B. Abbott and P. Shaw, "Virtual nursing avatars: Nurse roles and evolving concepts of care," *OJIN Online J. Issues Nurs.*, vol. 21, no. 3, 2016.

[93] "SENSELY." https://www.sensely.com/.

[94] M. Wu and J. Luo, "Wearable technology applications in healthcare: a literature review," *Online J Nurs Inf*, vol. 23, p. a, 2019.

[95] M. Awais, L. Palmerini, A. K. Bourke, E. A. F. Ihlen, J. L. Helbostad, and L. Chiari, "Performance evaluation of state of the art systems for physical activity classification of older subjects using inertial sensors in a real life scenario: A benchmark study," *Sensors*, vol. 16, no. 12, p. 2105, 2016.

[96] N. Pannurat, S. Thiemjarus, and E. Nantajeewarawat, "A hybrid temporal reasoning framework for fall monitoring," *IEEE Sens. J.*, vol. 17, no. 6, pp. 1749–1759, 2017.

[97] C.-Y. Hsieh, K.-C. Liu, C.-N. Huang, W.-C. Chu, and C.-T. Chan, "Novel hierarchical fall detection algorithm using a multiphase fall model," *Sensors*, vol. 17, no. 2, p. 307, 2017.

[98] R. M. Gibson, A. Amira, N. Ramzan, P. Casaseca-de-la-Higuera, and Z. Pervez, "Matching pursuit-based compressive sensing in a wearable biomedical accelerometer fall diagnosis device," *Biomed. Signal Process. Control*, vol. 33, pp. 96–108, 2017.

[99] I.-M. Lee *et al.*, "Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy," *Lancet*, vol. 380, no. 9838, pp. 219–229, 2012.

[100] H. A. Frank, K. Jacobs, and H. McLoone, "The effect of a wearable device prompting high school students aged 17-18 years to break up periods of prolonged sitting in class," *Work*, vol. 56, no. 3, pp. 475–482, 2017.

[101] E. E. Dooley, N. M. Golaszewski, and J. B. Bartholomew, "Estimating accuracy at exercise intensities: a comparative study of self-monitoring heart rate and physical activity wearable devices," *JMIR mHealth uHealth*, vol. 5, no. 3, p. e34, 2017.

[102] C. Robinson-Cohen *et al.*, "Assessment of physical activity in chronic kidney disease," *J. Ren. Nutr.*, vol. 23, no. 2, pp. 123–131, 2013.

[103] A. Rozanski, "Assessment and Management of Psychosocial Risk Factors Within Preventive Cardiology Practice," in *ASPC Manual of Preventive Cardiology*, Springer, 2021, pp. 61–72.

[104] C. Jaeschke *et al.*, "Overview on SNIFFPHONE: a portable device for disease diagnosis," in *2019 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*, 2019, pp. 1–2.

[105] M. M. H. Shandhi *et al.*, "Cardiac Function Monitoring for Patients Undergoing Cancer Treatments Using Wearable Seismocardiography: A Proof-of-Concept Study," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 4075–4078.

[106] G. Gresham *et al.*, "Wearable activity monitors to assess performance status and predict clinical outcomes in advanced cancer patients," *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1–8, 2018.

[107] A. Derungs, C. Schuster-Amft, and O. Amft, "Wearable motion sensors and digital biomarkers in stroke rehabilitation," *Curr. Dir. Biomed. Eng.*, vol. 6, no. 3, pp. 229–232, 2020.

[108] J. H. Burridge *et al.*, "Telehealth, wearable sensors, and the internet: will they improve stroke outcomes through increased intensity of therapy, motivation, and adherence to rehabilitation programs?," *J. Neurol. Phys. Ther.*, vol. 41, pp. S32–S38, 2017.

[109] C.-K. Tey, J. An, and W.-Y. Chung, "A novel remote rehabilitation system with the fusion of noninvasive wearable device and motion sensing for pulmonary patients," *Comput. Math. Methods Med.*, vol. 2017, 2017.

[110] P.-L. Chia and D. Foo, "Overview of implantable cardioverter defibrillator and cardiac resynchronisation therapy in heart failure management," *Singapore Med. J.*, vol. 57, no. 7, p. 354, 2016.

[111] H. U. Klein *et al.*, "Bridging a temporary high risk of sudden arrhythmic death. Experience with the wearable cardioverter defibrillator (WCD)," *Pacing Clin. Electrophysiol.*, vol. 33, no. 3, pp. 353–367, 2010.

[112] M. Hernandez-Silveira *et al.*, "Assessment of the feasibility of an ultra-low power, wireless digital patch for the continuous ambulatory monitoring of vital signs," *BMJ Open*, vol. 5, no. 5, p. e006606, 2015.

[113] R. R. Kroll, J. G. Boyd, and D. M. Maslove, "Accuracy of a wrist-worn wearable device for monitoring heart rates in hospital inpatients: a prospective observational study," *J. Med. Internet Res.*, vol. 18, no. 9, p. e253, 2016.

[114] E. M. Goldberg and P. D. Levy, "New approaches to evaluating and monitoring blood pressure," *Curr. Hypertens. Rep.*, vol. 18, no. 6, p. 49, 2016.

[115] G. Bilo, C. Zorzi, J. E. O. Munera, C. Torlasco, V. Giuli, and G. Parati, "Validation of the Somnotouch-NIBP noninvasive continuous blood pressure monitor according to the European Society of Hypertension International Protocol revision 2010," *Blood Press. Monit.*, vol. 20, no. 5, p. 291, 2015.

[116] P. Muntner *et al.*, "Measurement of blood pressure in humans: a scientific statement from the American Heart Association," *Hypertension*, vol. 73, no. 5, pp. e35–e66, 2019.

[117] T. Fujikawa, O. Tochikubo, N. Kura, T. Kiyokura, J. Shimada, and S. Umemura, "Measurement of hemodynamics during postural changes using a new wearable cephalic laser blood flowmeter," *Circ. J.*, vol. 73, no. 10, pp. 1950–1955, 2009.

[118] T. Hampton, "Fully automated artificial pancreas finally within reach," *Jama*, vol. 311, no. 22, pp. 2260–2261, 2014.

[119] J. Wang, "Electrochemical glucose biosensors," *Chem. Rev.*, vol. 108, no. 2, pp. 814–825, 2008.

[120] "Glutrac." https://www.add-care.net/.

[121] P. Jain, A. M. Joshi, and S. P. Mohanty, "iGLU: an intelligent device for accurate noninvasive blood glucose-level monitoring in smart healthcare," *IEEE Consum. Electron. Mag.*, vol. 9, no. 1, pp. 35–42, 2019.

[122] "AERBETIC." https://www.aerbetic.com/.

[123] H. Lee *et al.*, "A graphene-based electrochemical device with thermoresponsive microneedles for diabetes monitoring and therapy," *Nat. Nanotechnol.*, vol. 11, no. 6, pp. 566–572, 2016.

[124] J. G. Nutt and G. F. Wooten, "Diagnosis and initial management of Parkinson's disease," *N. Engl. J. Med.*, vol. 353, no. 10, pp. 1021–1027, 2005.

[125] Z. Lin, H. Dai, Y. Xiong, X. Xia, and S.-J. Horng, "Quantification assessment of bradykinesia in Parkinson's disease based on a wearable device," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 803–806.

[126] M. Delrobaei, N. Baktash, G. Gilmore, K. McIsaac, and M. Jog, "Using wearable technology to generate objective Parkinson's disease dyskinesia severity score: Possibilities for home monitoring," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1853–1863, 2017.

[127] A. Marcante *et al.*, "Foot Pressure Wearable Sensors for Freezing of Gait Detection in Parkinson's Disease," *Sensors*, vol. 21, no. 1, p. 128, 2021.

[128] A. P. Allen *et al.*, "A systematic review of the psychobiological burden of informal caregiving for patients with dementia: Focus on cognitive and biological markers of chronic stress," *Neurosci. Biobehav. Rev.*, vol. 73, pp. 123–164, 2017.

[129] J. Kim *et al.*, "Non-invasive mouthguard biosensor for continuous salivary monitoring of metabolites," *Analyst*, vol. 139, no. 7, pp. 1632–1636, 2014.

[130] T. Salafi and J. C. Y. Kah, "Design of unobtrusive wearable mental stress monitoring device using physiological sensor," in *7th WACBE World Congress on Bioengineering 2015*, 2015, pp. 11–14.

[131] J. Zhu, L. Ji, and C. Liu, "Heart rate variability monitoring for emotion and disorders of emotion," *Physiol. Meas.*, vol. 40, no. 6, p. 64004, 2019.

[132] S. Murali, F. Rincon, and D. Atienza, "A wearable device for physical and emotional health monitoring," in *2015 Computing in Cardiology Conference (CinC)*, 2015, pp. 121–124.

[133] A. Steptoe and M. Marmot, "Impaired cardiovascular recovery following stress predicts 3-year increases in blood pressure," *J. Hypertens.*, vol. 23, no. 3, pp. 529–536, 2005.

[134] J. W. Matiko *et al.*, "Wearable EEG headband using printed electrodes and powered by energy harvesting for emotion monitoring in ambient assisted living," *Smart Mater. Struct.*, vol. 24, no. 12, p. 125028, 2015.

[135] J. Perdiz, G. Pires, and U. J. Nunes, "Emotional state detection based on EMG and EOG biosignals: A short survey," in *2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG)*, 2017, pp. 1–4.

[136] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE Trans. Consum. Electron.*, vol. 64, no. 2, pp. 196–203, 2018.

[137] R. Martinez, A. Salazar-Ramirez, A. Arruti, E. Irigoyen, J. I. Martin, and J. Muguerza, "A self-paced relaxation response detection system based on galvanic skin response analysis," *IEEE Access*, vol. 7, pp. 43730–43741, 2019.

[138] T. Kageyama, N. Nishikido, T. Kobayashi, Y. Kurokawa, T. Kaneko, and M. Kabuto, "Self-reported sleep quality, job stress, and daytime autonomic activities assessed in terms of short-term heart rate variability among male white-collar workers," *Ind. Health*, vol. 36, no. 3, pp. 263–272, 1998.

[139] R. Bin Hossain, M. Sadat, and H. Mahmud, "Recognition of human affection in smartphone perspective based on accelerometer and user's sitting position," in *2014 17th International Conference on Computer and Information Technology (ICCIT)*, 2014, pp. 87–91.

[140] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "Moodscope: Building a mood sensor from smartphone usage patterns," in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 2013, pp. 389–402.

[141] W. Hsu, C. Baumgartner, and T. M. Deserno, "Notable Papers and Trends from 2019 in Sensors, Signals, and Imaging Informatics," *Yearb. Med. Inform.*, vol. 29, no. 1, p. 139, 2020.

[142] B. S. Chandra, C. S. Sastry, and S. Jana, "Robust heartbeat detection from multimodal data via CNN-

based generalizable information fusion," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 3, pp. 710–717, 2018.

[143] Y. Yuan and K. Jia, "FusionAtt: deep fusional attention networks for multi-channel biomedical signals," *Sensors*, vol. 19, no. 11, p. 2429, 2019.

[144] T. Zhu, M. A. F. Pimentel, G. D. Clifford, and D. A. Clifton, "Unsupervised bayesian inference to fuse biosignal sensory estimates for personalizing care," *IEEE J. Biomed. Heal. informatics*, vol. 23, no. 1, pp. 47–58, 2018.

[145] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nat. Med.*, vol. 25, no. 8, pp. 1301–1309, 2019.

[146] O. Yildirim, U. B. Baloglu, R.-S. Tan, E. J. Ciaccio, and U. R. Acharya, "A new approach for arrhythmia classification using deep coded features and LSTM networks," *Comput. Methods Programs Biomed.*, vol. 176, pp. 121–133, 2019.

[147] Q. Shi *et al.*, "Progress in wearable electronics/photonics—Moving toward the era of artificial intelligence and internet of things," *InfoMat*, vol. 2, no. 6, pp. 1131–1162, 2020.

5. The State of the Art in Public Health Informatics

Public health informatics is defined as the systematic application of information, computer science, and technology to public health practice, research, and learning. Health informatics brings state-of-the-art technology in the healthcare sector and specifically the idea of a "smart care plan", which is an important factor for paving the way toward Predictive, Preventive, Personalized and Participatory Medicine (P4-medicine). Moreover, the HL7 FHIR standard and blockchain ledger technologies are two very promising areas of research and development for health information management. An appropriate interaction among their approaches, concepts and tools could give rise to the hub of the IT infrastructure for P4-medicine. This overview presents current issues and proposed solutions in health domain, in the form of tools and software applications to support data collection, analysis and recording, advanced IT systems such as blockchain to manage the different phases of health processes, standards to enable fast healthcare interoperability, and the adoption of electronic health records and clinical pathways to foster P4-medicine. Finally, topics of privacy issues and other challenges, as well as opportunities of applying health informatics are discussed.

5.1 Needs and Tools for the Health Domain

The total amount of health expenses represents the amount spent on health care and related activities such as administration of insurance, health research, and public health, including expenses from both public and private funds. By 2000, health spending reached about $1.4 trillion in the USA, and in 2019 the amount spent was doubled to $3.8 trillion [1]. This emerging trend reflects a gap in the healthcare sector where the "value-based" policies of commercial and government insurance companies try to fill by shifting the attention of clinical organizations from individual patient visits to managing larger populations and improving their overall health while being cost effective. A number of health information technology (IT) solutions, such as health information exchanges (HIEs), have urged the collaboration of health systems and public health departments to better manage their overlapping "community" denominators to integrate across many different digital silos [2].

The major innovations that must be provided to the health system should be able to provide [3]:

- A universal model for health focused on the individual person (each time, and not only for a specific clinical event)
- A proactive approach to the health sector, using novel tools aiming at including the patient in the care processes.
- An integrated process management, by generating cooperative care models through the digital connection among all the actors included in the prevention, treatment, and follow-up processes.
- A certification of the health protocols along with the clinical data produced, aiming to encourage a native use of knowledge technologies that allow to offer intelligent services capable of integrating and configuring themselves with a view to socio-health care comprehension as a complex adaptive system.

To achieve the above tasks, the health domain needs innovative IT platforms and services that comply with the most health informatics standards, capable of supporting stakeholders in the development of innovative, certified, and interoperable eHealth applications. Table 10 outlines current specific issues in healthcare and proposed solutions offered by IT platforms and services.

Table 9. Possible solutions for the main health issues [3].

Issue	Solution
Secure sharing of health data	Consolidated interoperability models and secure protocols must be adopted for exchanging heterogeneous information coming from different sources (like hospitals, first aids, laboratories, etc.), assuring privacy maintenance.
Personalized health care	Specific IT systems should be designed to gather, process, and store patient health information in a certified manner directly in the patient's home.
Health processes	Advanced models and tools for the optimization, certification, and handling of the health processes are necessary support for the decision-making phase.
Evidence-based medicine	An information model should be used to integrate and analyze large amounts of socio-health data, based on the adoption of the Big Data Analytics paradigm.
Internet of Things	Practical tools should be developed to effectively integrate the data produced by the numerous existing biomedical sensors and wearable devices with other patient-related data.

5.1.1 Widely used tools and applications

In the following, table 11 presents different tools for supporting survey and data collection, table 12 presents tools for analysis, visualization and reporting (AVR) of the generated data, and table 13 presents current systems that have been employed, which make use of social media information in order to support public health informatics (e.g., for active case finding, contact identification and prevention messaging in case of an infection outbreak).

Table 10: Support of Survey and Questionnaire Data Collection [4].

Application	Comments
Outbreak management components of reportable disease surveillance systems.	• Available as an integrated management component through the public health agency's reportable disease surveillance system [5]. • Availably commercial off-the-shelf products (e.g., Maven [The Apache Software Foundation, Wakefield, MA; https://maven.apache.org/ external icon]) or health department– designed and developed (e.g., Florida Department of Health's Merlin system).
Epi Info (CDC, Atlanta, GA)	• Free public-domain suite of software tools designed and maintained by CDC for public health practitioners and researchers. • Easy to set up; can be utilized to support mobile data collection. • Contains customizable data entry forms and database construction. • Enables data analyses with epidemiologic statistics, maps, and graphs for public health professionals who lack an IT background.

	• Used in outbreak investigations and for developing small-to-mid sized disease surveillance systems. • Useful for public health field investigators to know and use because of its capabilities. • Available for free download at http://www.cdc.gov/epiinfo
REDCap (Vanderbilt University, Nashville, TN)	• Secure Internet application for building and managing online surveys and databases. • Used to collect virtually any type of data, including in environments compliant with electronic records legislation (21 Code of Federal Regulations Part 11), the Federal Information Security Management Act of 2002 (44 U.S. Code §3541), and the Health Insurance Portability and Accountability Act of 1996 (Public Law 104– 191, 110 Stat 1936). • Specifically designed to support online or offline data capture for research studies and operations. • Accessible through computers, tablets, and smartphones. • Available at no charge to not-for-profit institutions that join the REDCap Consortium at http://www.project-redcap.org

Table 11: Applications for Analysis, Visualization, and Reporting (AVR) [4].

Application	Comments
SAS (Statistical Analysis System; SAS Institute, Inc., Cary, NC)	• Statistical analysis software suite for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics. • Highly powerful software application. • Additional information available at https://www.sas.com
SPSS (IBM Corporation, Armonk, NY)	• Analytic software widely used in social science studies. • In addition to statistical analysis, features data management (e.g., selecting cases, reshaping files, or creating derived data) and data documentation. • Additional information available at http://www.ibm.com
ArcGIS (Esri, Redlands, CA)	• Designed to store, manipulate, analyze, manage spatial or geographic data. • Additional information is available at http://www.esri.com
R (R Foundation, Vienna, Austria)	• Free, open-source statistical analysis software. • Contains graphics capability and run programs stored in script files. • Associated with RStudio, an integrated development environment for R. • Additional information is available at http://www.rstudio.com
ESSENCE (Electronic Surveillance System for the Early Notification of Community-based Epidemics)	• Syndromic surveillance system operational in many jurisdictions and nationally as part of CDC's National Syndromic Surveillance Program. • Jurisdictional versions have different features or data sets. • Developed by the Johns Hopkins University Applied Physics Laboratory. • Enhancements developed through a collaboration among CDC, state and local health departments, and the Applied Physics Laboratory. • Additional information about the National Syndromic Surveillance Program and ESSENCE available at https://www.cdc.gov/nssp
SaTScan (Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA)	• Analyzes spatial, temporal, and space-time data by using scan statistics. • Available for free download at https://www.satscan.org/
BioMosaic (CDC, Atlanta, GA)	• Analytic tool that integrates demography, migration, and health data. • Available to designated CDC staff only.

	• Combines information about travel, disease patterns, and location of U.S. settlement of persons from other countries. • Combines complex data from multiple sources into a visual format, including maps and other types of graphics. • Developed through a collaboration in 2011 among CDC's Division of Global Migration and Quarantine, Harvard University, and the University of Toronto.
HealthMap (Boston Children's Hospital, Boston, MA)	• Free mapping utility. • Uses informal Internet sources (e.g., online news aggregators, expert-curated discussions, and validated official reports) for disease outbreak monitoring and real-time surveillance of emerging public health threats to achieve a unified and comprehensive view of the current global state of infectious diseases. • Available for use at http://www.healthmap.org

Table 12. Emerging crowdsourcing tools and applications [1].

Application	Comments
Mobile devices and APPs	• Mobile devices and multiple application tools can assist field investigators with public health surveillance
Single-use online forms	• SurveyMonkey (San Mateo, CA)
EpiCollect (Imperial College London, UK)	• Internet and mobile app for generating forms (e.g., questionnaires) and freely hosted project online sites for data collection • Data collected, including global positioning systems and media, by using multiple telephones • All data centrally viewable by using Google Maps, tables, or charts • Available for free download at http://www.epicollect.net/
Social media	• Yelp (Yelp, Inc., San Francisco, CA), Twitter (Twitter, Inc., San Francisco, CA), and Facebook (Facebook, Inc., Menlo Park, CA)

5.1.2 Blockchain

The blockchain technology designed in 2008 as the core data and programming structure of the Bitcoin cryptocurrency, has widely spread and evolved in the last two decades. In a blockchain network, any transaction task concerns endpoints that are authenticated through public keys of a given digital signature scheme, and the blockchain ledger consists of a continuously growing list of transaction records that are grouped in blocks, where each block contains a cryptographic hash of the previous block. Considering that a given block cannot be changed, all the previous blocks in the chain— with high probability—cannot be changed as well, due to the properties of the hash function. More particularly, if the last block in the chain is supposed to be uniquely generated, then these properties are inherited with high probability by all the other blocks, and the overall blockchain satisfies both the consistency and integrity properties. Current state-of-the-art blockchain technologies mainly recognize two different types: permission less and permissioned [6], and operate on different protocols, such as byzantine fault tolerant (BFT) [7], proof of work (PoW) [8], proof of stake (PoS) [9], proof of activity (PoA) [10], proof of elapsed time (PoET) [11], etc.

Blockchain technology allows to implement innovative platforms in the health sector, facilitating the management of the different phases of the health processes, detecting, and certifying activities and procedures to be followed. This will facilitate the resources to be used, to monitor and optimize overall efficiency and effectiveness. Moreover, they will simplify the activities of medical and health personnel, offering to patients a better and faster treatment service. The certification of clinical data produced and

health processes performed will permit to provide "controlled" intelligent services to doctors in both: (i) the management of decision-making processes carried out in diagnostic, therapeutic and rehabilitation practice, and (ii) the assessment of the interventions to be carried out to provide patient health care. Indeed, this would allow training artificial intelligence-based systems on correct, verified information and improve the overall quality of services, reducing the health risk and ensuring alignment with clinical guidelines.

5.1.3 Electronic and Personal Health Records

Electronic health records (EHRs) are real-time records that make information available securely to authorized users. While an EHR contains the medical and treatment histories of patients, an EHR system is built to go beyond standard clinical data collected in a provider's office and can be inclusive of a broader view of a patient's care. EHRs are a vital part of health IT and are able to:

1. Involve a patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, laboratory, and test results.
2. Allow access to evidence-based tools that providers can use to make decisions about a patient's care.
3. Automate and streamline provider workflow.

One of the key features of an EHR is that health information can be created and managed by authorized providers in a digital format capable of being shared with other providers from different health care organization. EHRs are built to share information among different health care providers and organizations, such as laboratories, specialists, medical imaging facilities, pharmacies, emergency facilities, and workplace clinics, so they contain information from all clinicians involved in a patient's care. Many efforts have been performed worldwide to realize distributed EHR systems, even if with several critical issues. The implementation of these systems can be completed only with the deployment of numerous subsystems by many different actors (hospitals, clinical laboratories, general practitioner ambulatories, etc.) and by preserving the user privacy.

Furthermore, a Personal Health Record (PHR) is used to collect personal health information from the patient, like clinical reports, annotations or data produced by biomedical sensors. They represent an important tool complementary to EHRs, considering their ability to classify and memorize all the data provided by a patient, offering an individual's medical history. The main difference between EHR and PHR lies in the nature of the health information collected. EHRs gather certified clinical information produced by healthcare facilities, while PHRs gather information obtained by the patients and, for this reason, these data are not certified.

The volume produced by EHRs and PHRs may be large, thus the review of such data will be time-consuming and labor intensive. However, if the EHRs and PHRs follow specific standards for the way they are collected and have complete structured forms instead of containing free context, then automation algorithms will be enabled to extract and process information from such data sources. A systematic review of automated information extraction from EHRs for the detection of different diseases can be found in [12].

5.2 Standards

Many health informatics standards have been produced by the Standard Developing Organizations (SDOs) to ensure homogeneous implementation and interoperability of health IT systems. Standards are used to implement health record and workflow systems. In addition, standards have been used and integrated with all the new technologies to implement additional IT applications. The most current health

informatics standards and technical specifications, which refer to clinical data representation and care planning are the HL7 FHIR and IHE PCC DCP.

5.2.1 Fast Healthcare Interoperability Resources

Fast Healthcare Interoperability Resources (FHIR) is a new generation standards framework employed by Health Level Seven (HL7) International, which provides interoperability specification for the exchange of electronically healthcare information. The major goal of FHIR is to simplify the implementation of health IT applications, without sacrificing information integrity. It offers a consistent and easy to implement mechanism for exchanging data between healthcare applications.

The FHIR fundamental principle is the expression of the following key points: (i) focus on developers; (ii) support for common scenarios; (iii) leverage web technologies; (iv) human readability as a basis for interoperability; (v) content available for free. Regarding the data transparency, it acts as an 'open API' to access the data present in the numerous EHR systems. Also, regarding analytics, FHIR utilizes data structures that permit to decompose information for data analysis.

FHIR provides many improvements over existing standards, in particular: (i) focus on implementation; (ii) multiple implementation libraries with several examples; (iii) the specification is free; (iv) interoperability out-of-the-box—base resources can be used, but can also be adapted for local requirements; (v) evolutionary development path from HL7 v2 and CDA—standards can co-exist and leverage each other; (vi) based on web standards like XML, JSON, HTTP, Atom, OAuth, etc.; (vii) support for RESTful architectures and seamless exchange of information using messages or documents; (viii) comprehensive and easily understandable specifications; ix) relied on a human-readable format to be easily used by developers; (x) consolidated ontology-based analysis with a stringent formal mapping for correctness.

The current version of the FHIR specifications is 4.0.1, available on the website HL7 FHIR 2020. The specifications are categorized into several levels; each of them involves a particular aspect of the standard. More specifically, level 1 supports the overall infrastructure of the FHIR specification, preserving the main documentation for the FHIR specification. Level 2 is responsible for the implementation and binding to external specifications. Level 3 connects real-world concepts in the healthcare system. Level 4 offers resources to record and exchange data for the healthcare process and level 5 provides the ability to rationalize about the healthcare process. The major concepts of the FHIR standard are described below [3].

Resources

A resource is a main building block that can be utilized to store and exchange data to manage healthcare information and processes. A resource includes a set of structured data items and a human-readable XHTML representation of its content. Resources are collected in the following classes:

- Administration: covers basic data that can be represented in FHIR, such as Patient, Practitioner, Care Team, etc.
- Clinical: includes clinical records (e.g. Allergy, Procedure, Care Plan/Goal, ServiceRequest)
- Diagnostics: holds clinical diagnostics, including laboratory tests, imaging, and genomics
- Medication: contains the ordering, dispensing, administration of medications
- Workflow: involves the resources for managing assistance processes (e.g. appointment, order, encounter, etc.)
- Financial: supports billings and payments
- Clinical Reasoning: allows to provide the ability to reason, such as clinical decision support rules, quality measures, etc.

Data Types

The data types are utilized to categorize the resource elements. They are divided into the following four categories:

- Simple/primitive types, which are single elements with a primitive value;
- General-purpose complex types, which are re-usable clusters of elements;
- Metadata types, which are a set of types used with metadata resources;
- Special purpose data types, which are defined elsewhere in the specification for specific usages.

Bundling

Bundling is called the operation executed on resources to collect them into a single instance, including correlated data with respect to a specific context.

Profile

A FHIR profile is a set of rules that allow a FHIR resource to include specific constraints or extensions, so that further attributes can be added.

5.2.2 IHE PCC DCP

The Healthcare Enterprise (IHE) is an international organization promoted by healthcare professionals and industries with the purpose of improving the way computer systems in healthcare share information by using consolidated standards [13]. IHE is organized by clinical and operational domains, where interoperability and issues related to clinical workflows, information sharing and improved patient care in the areas of healthcare are determined.

IHE is based on a process in which devoted groups collect case requirements, define standards, and develop technical specifications. The documents produced, named Integration Profiles, specify how actors use standards to address a specific healthcare use case, by exchanging a set of structured messages named transactions. In IHE a transaction is defined as an interaction between actors that transfers the required information through standards-based messages.

The Integration Profiles are published by each IHE domain as part of their technical frameworks. The publication process is divided in different states [14]:

- Final Text (FT): stable;
- Trial Implementation (TI): frozen for trial use; changes allowed prior to FT;
- Public Comment (PC): a TI profile republished, or a new profile published for receiving public comments;
- Draft Supplement: not yet ready for Public Comment;
- Deprecated/Retired: no longer suggested or maintained by IHE

Vendors can assess the compliance of their implementations of Integration Profiles with the technical specifications during periodical events named IHE Connectathons, which offer a detailed implementation and testing process. These events are organized annually by the Associations affiliated to IHE International, which are IHE Europe, IHE North America, IHE South America, IHE Asia-Oceania, IHE Middle East.

General clinical care aspects such as document exchange, order processing, workflows and coordination with other specialty domains are dealt within the IHE PCC domain, sponsored by HIMSS (Health Information Management Systems Society) and ACP (American College of Physicians). Some solutions to these issues have been described in numerous Integration Profiles [15].

Specifically, the structures and transactions for care planning and sharing Care Plans that meet the needs of interested users are provided in the Dynamic Care Planning (DCP) Integration Profile, whose Revision 3.1 was published in September 2019 as Trial Implementation [16].

The DCP profile allows to dynamically update Care Plans by the different actors involved in the care processes. The profile takes advantage of these standards:

- From a functional point of view, it is based on HL7 Service Functional Model: Coordination of Care Service (CCS) [17].
- Regarding the data model, it derives its concepts from the HL7 Care Plan Domain Analysis Model (DAM) [17].
- Regarding the technical aspects, the profile is based on HL7 FHIR Resources and transactions.

The data that a system compliant to IHE PCC DCP should be able to process and to be represented in the following HL7 FHIR resources:

- Care Plan: tool used by clinicians to plan and manage care for an individual patient
- Plan Definition: an action definition that describes an activity to be performed
- Activity Definition: specific actions to be performed as part of care planning.

The actors formalized in this profile are described below:

- Care Plan Contributor: reads, creates and updates Care Plans and Plan Definitions, generates Care Plans and requests resources relied on a selected activity definition;
- Care Plan Service: coordinates Care Plans received from Care Plan Contributors and provides updated Care Plans to subscribed Care Plan Contributors;
- Care Plan Definition Service: coordinates Plan Definitions received from Care Plan Contributors and provides updated Plan Definitions to subscribed Care Plan Contributors;
- Care Team Contributor: reads, creates and updates Care Teams;
- Care Team Service: manages Care Teams received from Care Team Contributors and provides notification of updates and access to updated Care Teams to subscribers.

5.3 Clinical Pathways

Clinical Pathways (CPs) or clinical workflows are plans of care defined to implement the clinical guidelines. CPs are standardized descriptions of clinical processes for defined combinations of symptoms adapted to clinical conditions. They are tools that permit to outline, with regards to one or more pathologies or clinical problems, the best possible path within an organization and among organizations for taking care of the patient. CPs lie on the concept of putting a patient in a therapeutic diagnostic path where the medical team determines the most suitable therapy in agreement with the interested parties. The purpose of CPs should be to [3]:

- Include a clear explanation of the objectives and key elements of clinical healthcare based on scientific evidence;
- Make an easier communication among team members, caregivers and patients;

- Manage the healthcare processes by coordinating roles and implementing the activities of multidisciplinary teams;
- Include documentation, monitoring and evaluation of the outcomes;
- Identify the resources necessary to implement the path.
- Increase the quality of clinical care, improving outcomes and promoting patient safety through the use of the right necessary resources;
- Support health professionals, clinicians, and care operators, by improving the quality of services.

CPs are devised to support the professionals in this complex procedure: the design of the path, the execution, the evaluation of the different parameters that could lead to an improvement of the pathway and the possibility of managing the patient's conditions with respect to the specific identified needs. Furthermore, the integration and use of EHRs and other health information systems into CPs will permit the support of a multidisciplinary team of professionals from heterogeneous systems (hospitals, private clinics, etc.), improving the decision making process of physicians and the quality of patient care. Once the most suitable treatment path for the specific problem is defined, it is crucial that all health professionals and the patient follow the whole workflow. The ability to update the treatment plan is also essential to follow the specific needs of a patient during the start of the treatment or during the therapy, to set the treatment plan according to patient's needs. Nowadays, the more salient benefits of CP contain improved patient involvement in treatment procedures, reduced hospitalization times, improved overall medical quality, reduced medical costs, and reduced incidence of poor practices [18]. Fig. 9 demonstrates an example of a clinical pathway modeled in line with OMG BPMN 2.0[4] standard.

Clinical pathways are implemented in different medical domains. However, their application is typically difficult without an appropriate information communication technology (ICT) environment. It is very complex to implement the follow-up of the clinical process without a system able to support the physicians in using efficiently the collected data, performing actions, and analyzing results.

The definition of the IT services architecture based on informatics health standards (such as HL7, FHIR, IHE, etc.) and on the use of blockchain technologies will permit in a simple way to make the care plans: (i) interdisciplinary (among different departments and systems); (ii) connected to each other, and consequently allowing interaction among different actors with different roles, use different medical skills and allow the communication in an easy manner. In addition, a platform able to monitor all the phases of a clinical workflow would permit incentivizing the patient to take part in the treatment process: in this way, it would be possible to have the trust from the patient and thus increase the probability that he/she will follow the therapy correctly. Then, it would increase the degree of personalization of the clinical pathway in a secure way. The use of blockchain technology in the architecture of such a platform would contribute to the following important benefits:

- Identification of an integrated and verified treatment plan
- Management of care paths in a safe way, by satisfying confidentiality and integrity
- Log all the operations carried out on the clinical pathways for subsequent analysis phases, useful to certify the actions taken in the care process and possibly determine responsibilities in the procedure
- Guide physicians and patients to comply with the specific treatment plan
- Verification of the correct application of the CP specific to the situation: the system made up of blockchain technology can identify a deviation from the modeled CP and thus notify the observed deviation.

[4] This stands for Object Management Group, Business Process Model & Notation, URL: https://www.omg.org/spec/BPMN/2.0/PDF

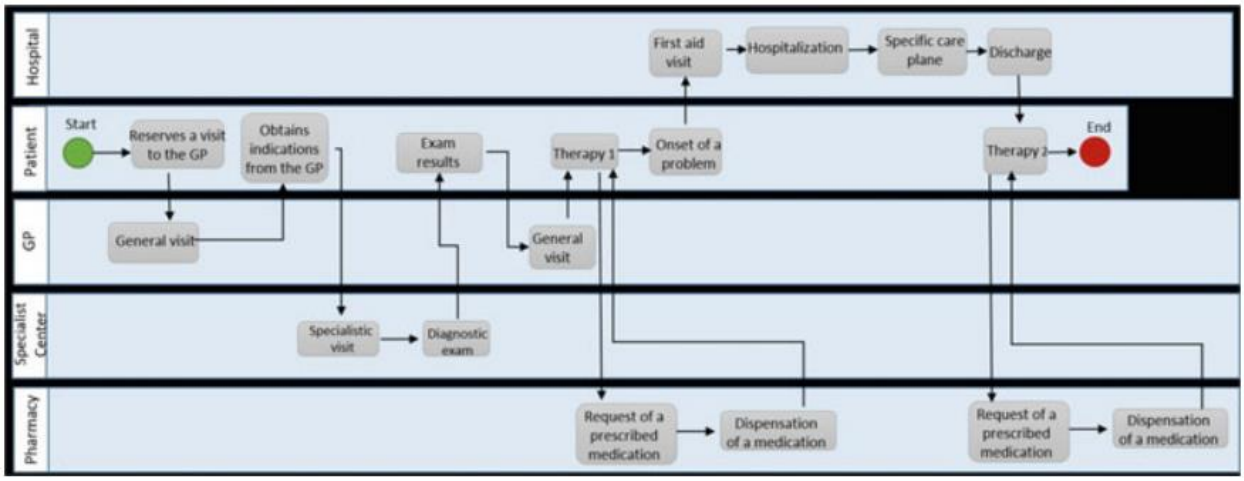

Figure 9: An example of a care plan [3].

5.4 Privacy Issues and Challenges for Health Informatics

Research on health informatics (HI) has many interesting challenges to face, in order for the HI sector to be transformed. Important challenging aspects are described below [19]:

1. **Conventional healthcare environment:** To embrace the full potential of the integration of the HI system into healthcare environments, conventional strategies, work-plans and use of medical devices must be ready for changes.

2. Infrastructure issue: HI brings new security problems and challenges. The issue of dynamic versus static network, with no fixed end focuses implies that a considerable amount of the existent communication mechanism for moving messages safely may not work properly. This implies that there are critical difficulties in the security and administration of this information, just as its assurance and security. At present, these associations are profoundly dependent on trust. If the medical device engineer does not consider cyber-attack threats while designing the devices, it can be termed as technical debt [20]. Technical debt may result in compromised medical devices with associated unpredictable behavior. The major concern is the impact of an exploitation rather than the exploitation itself.

3. Device diversity, interoperability, and vulnerability: Different types of medical wearable devices have been used in the health informatics system. So, interoperability between those devices is a big concern. It must be an interoperable system where data need to be transferred both one-to-one and one-to-many connections, including information exchange across multiple interfaces where the devices need to be compatible with one another. It is mandatory to consider that in any communication between multiple systems, the combination of interfaces is almost double. There is no collection of information regarding the capability of the devices. Device registry requires indexes of devices functionality, conventions, phrasings, and standards. The system's interoperability could make the system invulnerable to different threats.

4. Data integrity and consistency: Data integrity meaning of preserving the original data even in the case of any alterations. Ensuring integrity in a HI system guarantees the correctness of data which lead to minimizing errors and improving the safety of patients.

5. Privacy concern: Medical information is the most private kind of information and the access to this information is sensitive and must be approved by assigned experts [21]. Privacy in the healthcare sector may cause different results to the patients including the refusal of the administration to death. For specialist co-ops, privacy breaches can prompt legitimate authorizations, financial loss, or loss of goodwill. A comprehensive content analysis in security and privacy issues in e-health can be found [22].

6. Data access control: Information access to information control is used for ensuring privacy and security to any data. For HI, it is obligatory to control access across the whole system. The whole system is supposed to have different access segmentation. While guaranteeing CIA triads, protecting patients' vital pieces of information from unauthorized divulgence is fundamental under any conditions. Overall, the patient is characterized as the maker of the data. Building up the responsibility for data is important for securing the system from unauthorized access and manipulating the patient's health-related information. Authorization is mainly carried out by a security mechanism called access control. Medical data are stored in the cloud which is distributed covering a larger area. Sometimes it is a challenge for the system.

7. Human factors: Human factor plays the most crucial factor in HI and staff training is a prerequisite for deploying a technology-based health system. According to a study conducted by KTH university research students in Sweden over physicians, it was identified that around 76% of them considered human factor as the ultimate challenge in EHR implementation whereas 53% had almost no interest in Health IT [23]. Therefore, EHR systems have a higher probability of being successfully implemented if a usability study is carried out beforehand adopting to the healthcare environment.

8. Laws and ethics: Laws and ethics are the reason of privacy breach. Hospitals and governments provide records about the patient diseases to research agencies so that they could assist, e.g. in case of a disease outbreak. The government is supposed to assure that the research agencies deal with that information in the best way without causing any misuse of it.

9. Data authenticity: Authenticity is simply the validness of the data. Because of the lack of authentication of information, attacks like man-in-the-middle (MITM) could take place. To prevent this kind of attack, endpoint authentication is required in the cryptographic protocol.

10. Confidentiality and availability: The facility to protect the information in the HI system to be accessed by authorized subjects is called confidentiality. Confidentiality involves a lot of rules so that the private information could not be accessed by general mass. Authorized subjects receive access based on their working role. Thus, nothing about the patient's health record should be shared without their consent.

The system should have the option to be accessed whenever required by approved ones, for instance, on account of any crisis circumstance a particular doctor needs access to the patient's record to complete analysis and favor prescription to a patient. The system ought not to be obliged to a particular time generally; a patient may require a doctor's support any time.

11. Site recruitment: Site recruitment is the process of engaging a community health center and gaining their interest in participating in a collaborative network [24]. The key to site recruitment is that network must provide a mutually beneficial relationship between the site and other sites or users of the network. Site recruitment is usually hindered by financial and geographic barriers, regulatory issues when establishing business agreements between a new site and the shared network simply due to the sensitive nature of health data.

12. 4 V's: There are 4V's which are challenging for HI [25]. Those are volume, variety, velocity, and veracity. Volume refers to the exponential growth of medical data. Velocity refers to the need of real-

time processing for quick support and decision-making. Variety refers to the different nature of data coming from varied data sources. Finally, veracity refers to the proper data quality and reliability, since most of the times, data are often biased, full of clatter and anomalies that create a potential threat to proper decision-making processes and treatments for the patients.

5.5 Opportunities and Outcomes of Health Informatics

Health Informatics brings state-of-the-art technology in healthcare sectors. Both patients and clinicians are depending on new electronic technology and information systems.

Decision making support to improve patient care: For a patient, proper assessment/diagnostic and treatment is the most important of all. HI ameliorates the standard of treatment to patients by healthcare sector. HI helps data to be processed and recovered easily and effectively but can also be a resource of decision-making. Computerized protocols provide advantages that help make better decisions for physicians and clients.

Personalized treatment: HI will be able to provide personalized treatment to everyone in a very short time. It will improve the standard of treatment; patients will receive the best practice from the specialists, while doctors will be able to detect any diseases before a patient shows any symptoms. Furthermore, with the help of machine and deep learning techniques Recommender Systems (RS) can provide meaningful information to the patients depending on the specific requirements and availability of health records [26]. A recent study [27]reported the implementation of an Artificial Intelligence-Clinical Decision Support System (AI-CDSS) in 6 rural clinics in China, emphasizing the technical limitations and usability barriers, as well as issues related to transparency and trustworthiness of AI-CDSS.

Reduce treatment costs: Due to medical errors, every year a huge amount of money has been wasted. HI can reduce that cost at a larger amount. Connection and transparency between the partners in the healthcare sector will enable the adoption of “value-based” delivery models, where clinicians receive higher payment for delivering more efficient care.

Specifically, these models provide reimbursement for the improved health outcomes of a defined patient population rather than individual visits and services. This shift in the reimbursement models has motivated providers to better coordinate their entire patient population while controlling the overall cost of care. Under some of these models, value-based providers may receive a global budget for their assigned population, thus a reduced rate of utilization translates into larger shared-savings for both providers and payers. The latter has shifted the focus of value-based providers into prevention efforts (hence lower utilization) rather than costly treatment interventions, which fits well with the purposes of public health departments on reducing preventable diseases in large populations. In certain U.S. states such as Maryland, health care financing is substantially moving towards global budgets based on the size and characteristics of the population living in a catchment area, rather than based only on those who present themselves for services [28]. According to [29] a large component of an entire state’s health system’s budget is based on a population of people who may not utilize the health system at all. Data collected by public health departments are a key source of data required to calculate population-level health measures that will eventually be used to determine global budgets for medical care delivery systems and to evaluate whether they reach their community health goals, in order to achieve substantial financial incentives [30].

Study of chronic diseases: Data collected by a large health system or HIE may be used to study the prevalence of a specific chronic disease in a given geography [31], which is traditionally accomplished by

health departments using exhaustive survey methods for public health needs evaluation and monitoring purposes.

Remote monitoring: Caregivers do not require sitting the whole day by the side of patients to monitor their health status. HI will permit the clinicians to monitor the patients remotely and observe multiple patients at a time.

Telemedicine: Telemedicine is providing health services from remote distance through electronic signal [32]. HI is a part of telemedicine that covers distance healthcare service. Telemedicine can extend its dimension through HI as the use of computerized database, records, information access as well as decision making based on medical data.

Heredity analysis: HI can incorporate genome analysis in the conventional decision-making methods of health care by designing innovative and reliable tools for gene sequencing. This will generate a new way of public health treatment. The genomic advancements can facilitate inference, treating especially inherited infections and multi-faceted diseases [33].

5.5.1 Public Health Informatics 3.0

The implementation of Public Health Informatics 3.0 system may be the 'renaissance' of the above-mentioned opportunities. Health 3.0 is a health-related extension of the concept of Web 3.0 [34] whereby the users' interfaces with the data and information available on the web are personalized to enhance their experience. This is based on the concept of the Semantic Web [35], wherein websites' data is accessible for sorting to adjust the presentation of information based on user preferences. Health 3.0 will use such data access to enable individuals to better retrieve and contribute to personalized health-related information within networked electronic health records, and social networking resources [36,37]. Health 3.0 has also been described as the idea of semantically organizing electronic health records to create an Open Healthcare Information Architecture [38]. Health care could also make use of social media and incorporate virtual tools for improved interactions between health care providers and patients.

Health 3.0 aims to improve access to health-related information on the web via semantic and networked resources, to facilitate an improved understanding of health issues with the aim of increasing patient self-management, enhancing health professional expertise. In addition, Health 3.0 will foster the creation and maintenance of supportive virtual communities within which individuals can help each other to understand and manage common health-related issues. Furthermore, personalized social networking resources can also serve as a way for health professionals to enhance individuals' access to healthcare expertise, and to facilitate health professional-to-many-patients communication with the goal of improved acceptance, understanding and adherence to best therapeutic options.

However, many local communities face challenges implementing a Public Health 3.0 model. First and foremost, along with new informatics leadership roles for public health, there must be a workforce available to manage and analyze the data shared across the partners. It is essential for public health staff to have a thorough understanding of informatics, data flows, data collection processes, and the use of data at the public health agency in order to work productively with a larger and more diverse group of community partners. To be valuable to all partners, the electronic data available to the stakeholders must provide reliable and actionable data that will inform and add value to the component groups providing the data. This data sharing network will need to be transparent on how data are collated, protected, and shared with members of consortia.

So far, public health at a local level has been unable to integrate information technology. Furthermore, health departments face financial and resource shortages, specifically reduced government spending for public health.

5.6 References

[1] C. C. Rabah Kamal, Daniel McDermott, Giorlando Ramirez, "How has U.S. spending on healthcare changed over time?" https://www.healthsystemtracker.org/chart-collection/u-s-spending-healthcare-changed-time/#item-start.

[2] H. H. K. Kharrazi, D. Horrocks, and J. Weiner, "Use of HIEs for value-based care delivery: a case study of Maryland's HIE," in *Health Information Exchange: Navigating and Managing a Network of Health Information Systems*, Elsevier Inc., 2016, pp. 313–332.

[3] M. Ciampi, A. Esposito, F. Marangio, M. Sicuranza, and G. Schmid, "Modernizing Healthcare by Using Blockchain," in *Applications of Blockchain in Healthcare*, Springer, 2021, pp. 29–67.

[4] J. J. Hamilton and R. S. Hopkins, "Using technologies for data collection and management," *CDC F. Epidemiol. manual. CDC*, 2019.

[5] Public Health Informatics Institute (PHII), "Electronic Disease Surveillance System (EDSS) Vendor Analysis." [Online]. Available: https://www.phii.org/resources/view/4409/electronic-disease-surveillance-system-edss-vendor-analysis.

[6] "TOSHENDRA KUMAR SHARMA. PERMISSIONED AND PERMISSIONLESS BLOCKCHAINS: A COMPREHENSIVE GUIDE." [Online]. Available: https://www.blockchain-council.org/blockchain/permissioned-and-permissionless-blockchains-a-comprehensive-guide/.

[7] M. Pease, R. Shostak, and L. Lamport, "Reaching agreement in the presence of faults," *J. ACM*, vol. 27, no. 2, pp. 228–234, 1980.

[8] S. Nakamoto, "A peer-to-peer electronic cash system," [Online]. Available: https://www.klausnordby.com/bitcoin/Bitcoin_Whitepaper_Document_HD.pdf.

[9] S. King and S. Nadal, "Ppcoin: Peer-to-peer crypto-currency with proof-of-stake," *self-published Pap. August*, vol. 19, p. 1, 2012.

[10] I. Bentov, A. Gabizon, and A. Mizrahi, "Cryptocurrencies without proof of work," in *International conference on financial cryptography and data security*, 2016, pp. 142–157.

[11] L. Chen, L. Xu, N. Shah, Z. Gao, Y. Lu, and W. Shi, "On security analysis of proof-of-elapsed-time (poet)," in *International Symposium on Stabilization, Safety, and Security of Distributed Systems*, 2017, pp. 282–297.

[12] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, "Extracting information from the text of electronic medical records to improve case detection: a systematic review," *J. Am. Med. Informatics Assoc.*, vol. 23, no. 5, pp. 1007–1015, 2016.

[13] IHE International, "Making Healthcare Interoperable." https://www.ihe.net/.

[14] "Integrating the Healthcare Enterprise® (IHE)." https://wiki.ihe.net/index.php/Main_Page.

[15] "Patient Care Coordination," *International IHE*. https://www.ihe.net/ihe_domains/patient_care_coordination/.

[16] "IHE Patient Care Coordination Technical Framework Supplement." [Online]. Available: https://www.ihe.net/uploadedFiles/Documents/PCC/IHE_PCC_Suppl_DCP.pdf.

[17] "HL7 Service Functional Model; Coordination of Care Service (CCS), STU Release 1." [Online]. Available: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=452.

[18] G. Fico *et al.*, "Integration of personalized healthcare pathways in an ICT platform for diabetes managements: A small-scale exploratory study," *IEEE J. Biomed. Heal. informatics*, vol. 20, no. 1, pp. 29–38, 2014.

[19] M. H. Raju, M. U. Ahmed, and M. A. R. Ahad, "Health Informatics: Challenges and Opportunities," in *Signal Processing Techniques for Computational Health Informatics*, Springer, 2020, pp. 231–246.

[20] P. A. H. Williams and V. McCauley, "Always connected: The security challenges of the healthcare Internet of Things," in *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, 2016, pp. 30–35.

[21] J. Al-Muhtadi, B. Shahzad, K. Saleem, W. Jameel, and M. A. Orgun, "Cybersecurity and privacy issues for socially integrated mobile healthcare applications operating in a multi-cloud environment," *Health Informatics J.*, vol. 25, no. 2, pp. 315–329, 2019.

[22] N. A. Azeez and C. Van der Vyver, "Security and privacy issues in e-health cloud-based system: A comprehensive content analysis," *Egypt. Informatics J.*, vol. 20, no. 2, pp. 97–108, 2019.

[23] M. Kubbo, M. Jayabalan, and M. E. Rana, "Privacy and security challenges in cloud based electronic health record: towards access control model," in *The Third International Conference on Digital Security and Forensics (DigitalSec 2016)*, 2016, p. 113.

[24] D. R. Harris, T. J. Harper, D. W. Henderson, K. W. Henry, and J. C. Talbert, "Informatics-based challenges of building collaborative healthcare research and analysis networks from rural community health centers," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2016, pp. 513–516.

[25] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. S. Iyengar, "Computational health informatics in the big data age: a survey," *ACM Comput. Surv.*, vol. 49, no. 1, pp. 1–36, 2016.

[26] J. Saha, C. Chowdhury, and S. Biswas, "Review of machine learning and deep learning based recommender systems for health informatics," in *Deep Learning Techniques for Biomedical and Health Informatics*, Springer, 2020, pp. 101–126.

[27] D. Wang *et al.*, "' Brilliant AI Doctor' in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment," *arXiv Prepr. arXiv2101.01524*, 2021.

[28] J. M. Sharfstein, D. Kinzer, and J. M. Colmers, "An update on Maryland's all-payer approach to reforming the delivery of health care," *JAMA Intern. Med.*, vol. 175, no. 7, pp. 1083–1084, 2015.

[29] R. Gamache, H. Kharrazi, and J. P. Weiner, "Public and population health informatics: the bridging of big data to benefit communities," *Yearb. Med. Inform.*, vol. 27, no. 1, p. 199, 2018.

[30] E. Hatef, E. C. Lasser, H. H. K. Kharrazi, C. Perman, R. Montgomery, and J. P. Weiner, "A population health measurement framework: evidence-based metrics for assessing community-level population health in the global budget context," *Popul. Health Manag.*, vol. 21, no. 4, pp. 261–270, 2018.

[31] S. E. Perlman, K. H. McVeigh, L. E. Thorpe, L. Jacobson, C. M. Greene, and R. C. Gwynn, "Innovations in population health surveillance: using electronic health records for chronic disease surveillance,"

Am. J. Public Health, vol. 107, no. 6, pp. 853–857, 2017.

[32] K. Kasemsap, "The importance of telemedicine in global health care," in *Handbook of research on healthcare administration and management*, IGI Global, 2017, pp. 157–177.

[33] J. Parihar, P. Kansal, K. Singh, and H. Dhiman, "Assessment of bioinformatics and healthcare informatics," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*, 2019, pp. 465–467.

[34] J. Hendler, "Web 3.0 Emerging," *Computer (Long. Beach. Calif).*, vol. 42, no. 1, pp. 111–113, 2009.

[35] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Sci. Am.*, vol. 284, no. 5, pp. 34–43, 2001.

[36] K. B. DeSalvo, P. W. O'Carroll, D. Koo, J. M. Auerbach, and J. A. Monroe, "Public health 3.0: time for an upgrade," *Am. J. Public Health*, vol. 106, no. 4, p. 621, 2016.

[37] K. DeSalvo and Y. C. Wang, "Health informatics in the Public Health 3.0 era: intelligence for the chief health strategists," *J. Public Heal. Manag. Pract.*, vol. 22, no. Suppl 6, p. S1, 2016.

[38] CISION, "SemTech 2010 Speaks Out About Health 3.0-Open Healthcare Information Architecture." [Online]. Available:http://www.prweb.com/releases/2010/05/prweb3957314.htm