Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

AI and precision oncology in clinical cancer genomics: From prevention to targeted cancer therapies-an outcomes based patient care

Zodwa Dlamini [a,**], Amanda Skepu [a,b], Namkug Kim [a,c], Mahlori Mkhabele [a], Richard Khanyile [a,d], Thulo Molefi [a,d], Sikhumbuzo Mbatha [a,e], Botle Setlai [a,e], Thanyani Mulaudzi [a,e], Mzubanzi Mabongo [a,f], Meshack Bida [a,g], Minah Kgoebane-Maseko [a], Kgomotso Mathabe [a,h], Zarina Lockhat [i], Mahlatse Kgokolo [j], Nkhensani Chauke-Malinga [k], Serwalo Ramagaga [l], Rodney Hull [a,*]

[a] *SAMRC Precision Oncology Research Unit (PORU), SARChI Chair in Precision Oncology and Cancer Prevention (POCP), Pan African Cancer Research Institute (PACRI), University of Pretoria, Hatfield, 0028, South Africa*
[b] *Next Generation Health, Division 1, CSIR, Meiring Naude Road, Brummeria, Pretoria, 001, South Africa*
[c] *Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea*
[d] *Department of Medical Oncology, Faculty of Health Sciences, Steve Biko Academic Hospital, University of Pretoria, Hatfield, 0028, South Africa*
[e] *Department of Surgery, Faculty of Health Sciences, Steve Biko Academic Hospital, University of Pretoria, Hatfield, 0028, South Africa*
[f] *Department of Maxillofacial and Oral Surgery, School of Dentistry, University of Pretoria, Hatfield, 0028, South Africa*
[g] *Department of Anatomical Pathology, National Health Laboratory Service (NHLS), University of Pretoria, Hatfield, 0028, South Africa*
[h] *Department of Urology, Faculty of Health Sciences, Steve Biko Academic Hospital, University of Pretoria, Hatfield, Pretoria, 0028, UK*
[i] *Department of Radiology, Faculty of Health Sciences, Steve Biko Academic Hospital, University of Pretoria, Hatfield, 0028, South Africa*
[j] *Department of Dermatology, University of Pretoria, Steve Biko Hospital, Hatfield, 0028, South Africa*
[k] *Department of Plastic, Reconstructive and Aesthetic Surgery, Steve Biko Academic Hospital University of Pretoria, Hatfield, 0028, South Africa*
[l] *Department of Otorhinolaryngology, University of Pretoria, Steve Biko Academic Hospital, University of Pretoria, Hatfield, 0028, South Africa*

ABSTRACT

Precision medicine is the personalization of medicine to suit a specific group of people or even an individual patient, based on genetic or molecular profiling. This can be done using genomic, transcriptomic, epigenomic or proteomic information. Personalized medicine holds great promise, especially in cancer therapy and control, where precision oncology would allow medical practitioners to use this information to optimize the treatment of a patient. Personalized oncology for groups of individuals would also allow for the use of population group specific diagnostic or prognostic biomarkers. Additionally, this information can be used to track the progress of the disease or monitor the response of the patient to treatment. This can be used to establish the molecular basis for drug resistance and allow the targeting of the genes or pathways responsible for drug resistance. Personalized medicine requires the use of large data sets, which must be processed and analysed in order to identify the particular molecular patterns that can inform the decisions required for personalized care. However, the analysis of these large data sets is difficult and time consuming. This is further compounded by the increasing size of these datasets due to technologies such as next generation sequencing (NGS). These difficulties can be met through the use of artificial intelligence (AI) and machine learning (ML). These computational tools use specific neural networks, learning methods, decision making tools and algorithms to construct and improve on models for the analysis of different types of large data sets. These tools can also be used to answer specific questions. Artificial intelligence can also be used to predict the effects of genetic changes on protein structure and therefore function. This review will discuss the current state of the application of AI to omics data, specifically genomic data, and how this is applied to the development of personalized or precision medicine on the treatment of cancer.

1. Introduction

Precision medicine, otherwise known as personalized medicine, aims to treat patients through tailor-made therapies based upon the traits specific to the population group a patient belongs to or in ideal situations traits specific to that single patient. These specific traits often refer to the patient's genome, transcriptome or proteome, but can include other factors such as lifestyle,environment, and socio-economic status. Frequently, this involves sequencing the patient's genome, transcriptome or analyzing their proteome. A digital twin is a virtual copy of

* Corresponding author.
** Corresponding author.
E-mail addresses: Zodwa.Dlamini@up.ac.za (Z. Dlamini), Rodney.Hull@up.ac.za (R. Hull).

https://doi.org/10.1016/j.imu.2022.100965
Received 3 March 2022; Received in revised form 25 April 2022; Accepted 11 May 2022
Available online 14 May 2022

a real-world physical object. The more precise, detailed and recent the information describing an object is, would lead to a more detailed and accurate digital twin. In terms of precision medicine, this would mean a digital twin of a specific patient or a specific population group [1]. Artificial Intelligence (AI) can be defined as algorithms and computing frameworks that can be used to perform various tasks that would normally rely on human intelligence in the form of reasoning, decision-making, speech recognition, language understanding, and visual perception [2]. AI can simply be defined as software that attempts to emulate human thought processes to accomplish a task in the same manner as a human expert in the field [3]. Ultimately, the aim of AI in precision medicine is to identify patterns in data using models and algorithms that can then be used to make predictions. These predictions are initially performed then perfected by machine learning through the software's own learning algorithms [4].

AI relies on machines performing functions such as rule-based logic, machine learning (ML), deep learning (DL), natural language processing (NLP), and computer imaging [2]. The recent ability for technologies to generate large amounts of omics data, including genomic, transcriptomic, proteomic (phenotypic) and epigenomic data has contributed to the necessity of AI in the analysis of medical information. With respect to genomic and transcriptomic data, this increase is due to next generation sequencing (NGS) and for proteominc data this is due to the generation of large amounts of proteomic data using mass spectrometric analysis [2].

Genome-wide association studies (GWAS) have been responsible for generating vast amounts of genomic data and associating this data with specific diseases such as cancer. The use of this patient specific data in precision medicine relies on the accurate integration, analysis, and interpretation of this data to provide a complete overview of changes in gene expression profiles in a particular cancer patient (Fig. 1) [5–7]. The resulting analysis can show changes in metabolic and signaling pathways specific to the patient. In this way it can be used to personalize the response of health care professionals to a single patient or group of patients. This multidimensional approach offers many advantages over traditional single-layer analysis (analysis undertaken on a single feature) [8,9]. In order to do this AI must be taught to recognize features in data. For this review these features will generally refer to patterns in the genomes and transcriptomes associated with a particular disease, outcome, or treatment response. These patterns are initially learned through the process of machine learning, through the initial analysis of large datasets, teaching datasets, and human guided feature identification. Once the AI has learnt to sort and classify data depending on techniques it has learnt, it can then act independently to analyse other new large datasets. With the increase in computing power, machine learning (ML) has advanced to become deep learning (DL). This makes use of computers to construct neural networks, allowing for multiple tasks to be divided amongst the available computing power, resulting in more in depth analysis and greater automation [3]. Frameworks that integrate data analysis and network-based approaches are used to capture, analyse, and select important patterns and profiles present in vast amounts of omics data, using realistic assumptions. Generally, these networks accomplish predictions for various parameters using a prior-posterior Bayesian structure, which is a probability distribution made using expected outcomes or predictions [4].

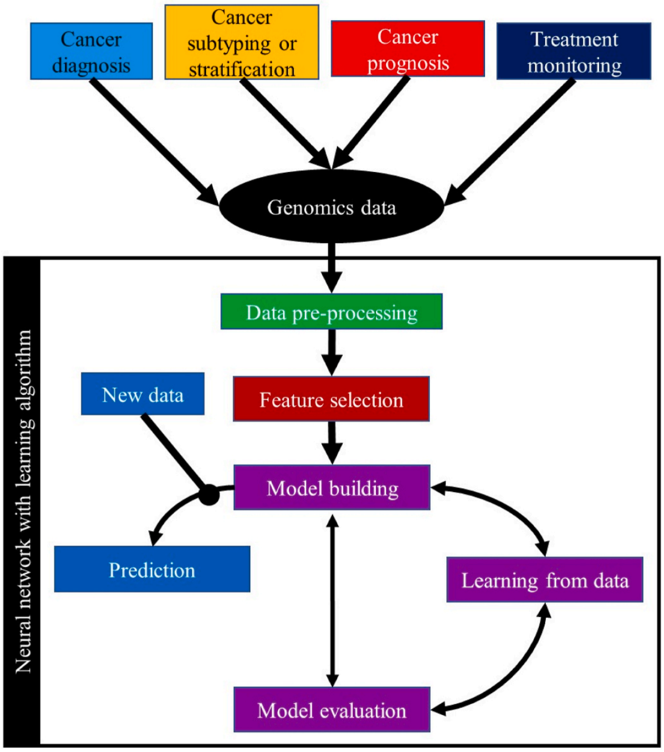

Fig. 1. General application of Artificial intelligence to genomics data: The purpose of the study dictates the type of data used. Problem definition and data selection is followed by data pre-processing and feature selection. The model is then built using this data, machine, and deep learning approaches. This model then uses learning methods. The model is tested to evaluate it. This is then used to fine tune and improve the model until it is satisfactory. The model is then implemented to analyse new data.

2. Large data

The generation of large amounts of omics data and the usefulness and availability of genetic information resulting NGS has led to the current big data era. The handling of large data has been assisted through the development of large capacity data storage, allowing for the storage and curation of large data sets. This was initially due to the development of large capacity hard drives and has been increased further by the development of cloud storage, cloud scaling and server networks. The integration of health records and patient genomic data would further expand the potential for the development of improved patient care [10]. Manual interpretation of genomics data becomes impossible if the data is too large (scalability) and different individuals may obtain different results analysing the same dataset (reproducibility). This is also time consuming, and this problem grows as technologies allowing for the faster and cheaper generation of larger datasets advance without the accompanying increase in people capable of analysing this data. AI can be used to search large datasets for specific patterns in the data. It can also then be used to divide data into discrete units based on specific parameters that make for easier, faster and more accurate analysis [2].

Next-generation sequencing (NGS) is capable of generating large datasets concerning the genomic, transcriptomic or epigenomic profiles of tumor cells [11]. Targeted sequencing has been extensively used for the detection of mutations or expression changes in multiple genes that are used as biomarkers in various cancers. Targeted sequencing is currently the preferred method of detecting changes in these genes due to its lower cost, high sensitivity, and higher coverage. However, NGS offers the promise of the identification or examination of large genomic rearrangements and mutations in tumors. This allows for the detection of changes in other non-targeted genes which may play an equally important role in cancer development and progression [12]. NGS has diverse applications including whole-genome sequencing, whole-exome sequencing, whole RNA sequencing, poly-A-minus RNA sequencing, target sequencing, and methylation sequencing [13]. Before the development of NGS, gene expression profiling was achieved using microarrays. Both of these technologies allow for the identification of upregulated or downregulated gene transcripts and these can be used to infer what signaling pathways are up or down regulated [14]. Gene and gene expression profiles obtained via NGS have already been used for risk prediction, disease diagnosis, and for the development of targeted therapies. The gene expression profile data can also be combined with outcome data in order to establish the relationship between gene expression profiles and the outcomes of disease or treatment, although this can only benefit future patients. In a similar manner the expression patterns can be combined with cancer staging and subtype data to allow for future diagnostics and staging using gene expression profiles [15, 16]. The first step in the use of NGS to generate omics data for cancer studies or diagnostics, is to ensure that data generation and reporting are performed at high enough standards to allow for proper analysis and subsequent application of the results [17]. This includes the use of the correct reference controls for proper validation and calibration. Genomic DNA reference material for NGS has been prepared by the Centres for Disease Control (CDC) and Prevention's Genetic Testing Reference Material Coordination Program (GeT-RM) [18]. However, since the analysis of NGS results relies on AI and machine learning, these standards need to be constantly updated as computer software and hardware change. These standards will also help to standardise the procedures across newly developed NGS platforms and different laboratories [18].

2.1. Reduction of data complexity

The combination of various omics data is very complex and in order to deal with these large data sets they are often simplified which may result in the loss of information. Data complexity can be defined as the number of input features or variables in a given dataset. These factors are known as dimensions and data complexity can be called data dimensionality [19], as such the most prevalent of these simplification techniques is known as dimensional reduction [20]. Dimensions refer to the attributes that describe a specific data point. Data reduction is important as dimensionality complicates predictive modelling. Dimensionality reduction converts higher dimension data into a dataset with less dimensions. Simplifying the data in this way reduces the amount of storage space needed, lowers training times, increases the speed of analysis, and removes redundant data. Dimensional reduction can be achieved through feature selection and feature extraction. Feature selection involves the selection of only relevant features while discarding those features deemed to be irrelevant [21]. Filter methods use various tests to filter data based on whether data sets are significantly different from each other, these include tests such as Correlation, Chi-Square Test and ANOVA. Wrapping methods use machine learning algorithms to evaluate the performance of this data. The performance of the data dictates whether these features are used or rejected. Finally embedded methods use model training methods to test the importance of each feature in the training process [21]. Feature extraction involves transforming the space containing many dimensions into a space with fewer dimensions. This method results in less data loss. Some common feature extraction techniques include principal component analysis (PCA), linear discriminant analysis, kernel PCA and quadratic discriminant analysis. PCA involves the selection of highly variant data. This data will be present in more classes and decreases the dimensionality of the data [21]. Techniques such as multiple co-inertia analysis and multiple factor analysis attempt to decrease data loss due to simplification by mapping the data to lower dimensional space. This can be done by removing data attributes so that the data can be plotted in fewer dimensions [20]. Neural networks can be organised into a framework that can be used to group and analyse this simplified data. This is achieved through the use of algorithms to analyse graphed data and multi-level Bayesian models (using probabilities to replace uncertainties). These can organise and interpret changes in the molecular composition in these samples and the level of these molecules and in some cases the modification of molecules and molecular interactions. These molecular changes can be detected using various omics approaches, using realistic estimations of one or more parameters to analyse this omics data [4]. Many new artificial intelligence networks and software have been developed to help integrate patient health records and data from analysis such as whole genome sequencing or transcriptome sequencing. One such tool named MEDomics, has the capacity to continuously learn based on multimodal health data inputs. It organises the data and assesses its quality, with the final aim of providing a more accurate prognosis for an individual patient. This AI was tested using records consisting of large amounts of collected data from many decades of patient care. Since this was old data the patient outcome was already known. It was able to use the data in these records to accurately provide prognoses for these patients [22] thereby proving the usefulness of this AI as a tool for patient diagnosis.

2.2. Workflow for germline variant discovery

The process of identifying differences in sequencing data obtained from NGS is known as variant calling (Fig. 2). These genomic variations that impact the phenotype, and may arise due to changes affecting expression, splicing, and amino acid sequence. These changes may result from single base changes, insertions, or deletions. This is done by aligning the raw sequencing data to a reference genome. The quality of the data is then improved by removing duplicates, insertions, and deletions (indels), re-alignment (through frame shifting one of the sequences) and base recalibration. These editing steps are followed by the removal of false positives. Often this needs to be done using speculation [23]. This time-consuming process is also prone to bias and errors. However, AI can be used to increase speed and accuracy using neural networks. For example, CNNs have been successfully used to pre analyse data and perform variant calling. This resulted in improved diagnosis in lung cancer [24]. The Cerebro analysis tool uses random forest-based ML to incorporate data pre-processing and variation calling into an analysis pipeline that results in the improved identification of tumor associated mutations [25]. Successful pre-analysis has also been achieved using Google's Deep Variant tool, which analyses the data as though it were an image and functions as image recognition software. Additionally, standardized workflow software has been developed. One of these, the open-source software, Variant calling workflow (OVarFlow) aims to automate the process while optimizing its reproducibility, as well as reducing the need for massive computing power [26]. Once variants have been identified they need to be classified and then annotated. Although functional studies performed *in vivo* or *in vitro* are the most desirable to determine the role a mutation plays in disease development and progression, AI is being investigated as a partial replacement for these studies. Some of these *in silico* tools include PolyPhen and SIFT [27,28], while more advanced ML based techniques are being developed that predict the effect of mutations on the secondary structures of proteins and compare this to native proteins using homology modelling [29, 30]. These approaches have included the use of neural networks, such as deep neural networks (DNNs) [31–37]. Alternately, changes in protein sequence can be compared and plotted on a decision tree based on sequence differences [38,39]. This method can be further improved by constructing multiple decision trees using the random forest technique [40–42]. Some of these *in-silico* methods attempt to determine if a mutation can cause cancer by determining the distribution of these mutations within the coding region of a protein. Non-random distributions are more likely to be associated with the development and progression of cancer [30]. However, despite these tools assisting the selection and classification of variants, it is important to validate these computational tools. This validation is performed using a set of guidelines known as the critical assessment of genome interpretation (CAGI). These guidelines consist of challenges that were formulated through the comparison of variants that are predicted to cause diseases with experimentally validated results [43]. As of 2021, the fifth edition of CAGI is the latest edition and consists of fourteen challenges. The variant selection made through the use of AI must meet these challenges by matching predictions outlined by CAGI [43]. The fifth edition places a major focus on challenges arising due to splicing, clinical genomics, complex disease datasets and missense variants [43].

One of the main challenges to any AI using ML or DL is the use of

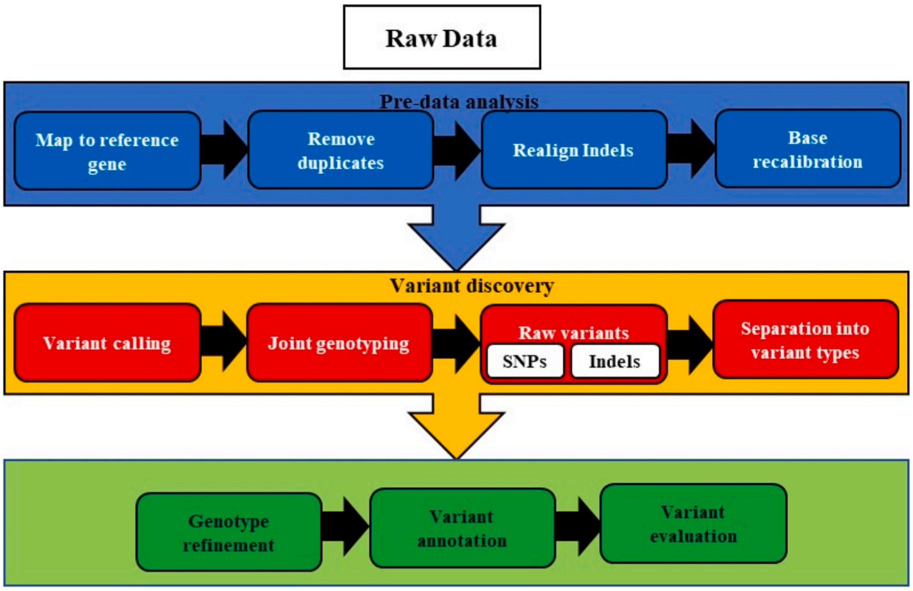

Fig. 2. Workflow for variant discovery. Data is first pre-processed by aligning it to a reference genome, removing duplicate sequences and correcting errors arising from insertions or deletions by frameshifting. Variants are then identified and classified and compared to known variants in the raw variant step. The resulting variants are sorted based on certain selection criteria. The accuracy of this data is improved in the refinement step before the variants are fully annotated and evaluated.

variances to classify cancer according to its pathogenicity and clinical relevance. Humans do this through a complex process. AI emulates this by using a custom set of rules. Ideally any AI would be able to learn these from a trained expert, following which it can apply and adapt these rules on its own. This set of rules must be comprised in such a way that it is able to adapt to a multitude of various scenarios. It has been reported that a logistic regression model is able to model a variable with a binary outcome using logarithmic functions to analyse probabilities. Using these models, AI was able to accurately predict patient situations and outcomes as well as treatment recommendations with a 1% false negativity and 2% false positivity rate. This is comparable to the predictions made by molecular pathologists [44].

A study was performed by Corti et al. (2019) in order to assess the ability of AI to provide prognostic and predictive information using NGS data of genes involved in colorectal cancer (CRC) [45]. Traditionally diagnosis and treatment decisions for CRC is performed by histopathology and Tumor Node Metastasis (TNM) classification. The AI they developed makes use of a worflow designed and executed using the IDEA® Data Analysis Software which uses NGS assays to detect and analyse CRC specific genomic target sequences. These sequences were 600 kb to 30 Mb in length. This workflow used decision making and learning algorithms to assess single nucleotide variants as described above. Clinically relevant molecular alterations were successfully identified and characterised [45]. These algorithms have also been used to analyse sequencing data obtained from circulating tumor DNA. This was successfully used to monitor changes in the progression of the disease, as well as to follow the response to treatment and establish if resistance to these treatments is developing [45]. This demonstrates the usefulness of developing and applying a variant calling worflow to be executed by AI in the treatment and monitoring of various cancers.

2.3. Other sources of large data

New sequencing related technologies are constantly being developed. Some of these new technologies include single cell sequencing and long read sequencing. Single cell sequencing allows for the genome or transcriptome of a single cell to be sequenced. This allows the sequence data to be context specific, by being associated with a single cell type, allowing for easier association between the changes in expression profiles and phenotypes and cellular functions. This method makes use of computational fluorescence microscopy and multiplex probe design to establish the expression of multiple genes in a single cell, and can also localise gene expression within the cell [46]. Long read sequencing is also known as third generation sequencing. It involves the sequencing of longer DNA sequences by focussing on a single molecule. This direct sequencing of a single molecule can be done in real time. This technology requires dedicated sequencing platforms such as the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore) platforms [47]. Long read sequences are also easier to assemble and can more accurately identify variants than short sequences and therefore would require less complex sequence assembly and analysis algorithms [47]. The usefulness of this technique in analysing single molecules or a single group of related molecules has been demonstrated in recent studies. The RAS family of proteins are well known for their role in carcinogenesis, being responsible for altered signalling pathways in a multitude of cancers, where they are commonly mutated or overexpressed. An examination of the splicing variants and multiple isoforms generated from the KRAS, NRAS and HRAS genes using long read sequencing, was able to analyse the splicing events that occurred. This gave information such as the exon/intron boundaries and 39 novel RAS mRNA transcript variants as well as the expression profile of these variants [48]. These splicing profiles could then be used as biomarkers for diagnosis, prognosis, the monitoring of the tumor response to treatment as well as identifying new targets for treatment.

The rapid evolution of cancer genomics has led to multiple meaningful evidence-based recommendations. In order to come to the right conclusion or recommendation it is important to make sense of vast amounts of available data sets. This includes the massive amounts of published literature relating to cancer genomics. Natural language processing (NLP) can be used by AI to select and extract specific genes, genetic variants, treatments, and conditions relating to cancer from the literature. These are named entities and the process of labelling individual entities is called Biomedical Named Entity Recognition (Bio-NER) [49]. The association of genetic alterations with recognized entities from the biomedical literature is necessary for the linking of data and an associated condition. One technique to do this involves the use of co-occurrence analysis. This involves two linked terms being assessed for the number of times they occur together, if this occurs more frequently than chance would dictate then it can be assumed that these terms are related to each other. This technique can give many false results (high recall and low precision) [50]. An example of the use of this technique is the study performed by de Ridder et al. 2007. Here the authors associated the terms “mutations related to insertions” and those mutations that are statistically more prevalent in cancers. These co-occurring terms were used to search the Retroviral Tagged Cancer Gene Database (RTCGD), This led to the identification of 86 mutations affecting cooperating oncogenes that function in tumorigenesist [51]. Another more accurate approach involves a role-based association, as performed by Hakenberg et al. 2012 who associated genetic variants with drug resistance and disease occuremce. These associations were used to mine abstracts in PubMed to identify SNPs (single nucleotide polymorphisms) involved in the drug resistant phenotype. The literature search identified 93% of the drug gene interactions found in the PharmGKB database. However, it is important to be specific indefining the search terms and roles to use to conduct this analysis. These can be difficult to choose and prone to bias [52]. The MEDscape algorithm uses NLP to analyses medical notes to update patient records and improve the accuracy of prognosis [22]. The final aim of any AI attempting to use large datasets to aid in the identification of biomarkers, is the association of these biomarkers with clinical endpoints. AI using NLP have been used to identify these association in patient medical records. Kehl et al. (2021) used AI performing NLP on 305 151 imaging reports and 233 517 oncologist notes from thousands of patients with multiple different tumor types. Their AI model extracted treatment outcomes from this data. These outcomes included cancer progression, treatment response and metastasis. Since these records were all old data the AI predictions correlated well with actual patient survival [53].

3. Neural networks, learning algorithms and decision-making tools

Machine learning (ML) is the term used to describe the ability of computers to learn to recognizes patterns in a large volume of data being used as training data [52]. This training process is used to create a mathematical model. Techniques used in ML include support vector machines, decision trees, factorization machines, logistic regression analysis and neural networks [54]. One of the most important factors when it comes to the use of ML is the scale and quality of training data [55]. Analysis performed by ML must be reproducible. This becomes difficult as the algorithms consist of many variables that can be set or tuned by the user [55]. The difference between ML and DL is that DL uses both supervised and unsupervised learning by integrating them using multi-layer non-linear analysis and classification. Deep learning is useful in automatically detecting image features and is therefore used in image classification [55], object detection and semantic segmentation. It is also used in AI applications such as natural language processing, and reinforcement learning [56,57].

The performance of a machine learning model must be validated. This must be done in an unbiased way using proper validation techniques. The choice of technique used depends on the situation [58]. The splitting of training data forms the basis of most validation techniques. Splitting the data and only showing the model one half of the training data allows the second set to be used as new data the model has never seen before. Its response to this data can then be assessed for model validation. The random split is normally performed in such a way that 70% is used for training while the remaining 30% is used as the test data [59]. The problem with this method arises due to different categories of data being present in a dataset. For example, if the data is age or sex specific. If the data split leads to one of these categories being over-represented, it can give rise to a sampling bias. Another problem may arise due to overfitting. This describes the situation where the test and training data result in the model being optimised for that dataset only and is “fixated” on the identification of parameters in a new dataset that were specific to the training dataset. To solve this a second split in the data can be made to create a holdout set, which is typically a 10% split. This data is then used to test the model a second time to ensure that overfitting has not occurred [59].

3.1. Neural networks

Neural networks attempt to imitate the way humans think, make decision and come to logical conclusions [60]. Neural networks use multiple neurons (in the form of fundamental computing units) to convert data from raw input data to classified, annotated, and analysed output data. Each neuron or node of the network applies weight to the input data and as such adds bias to the data. The node analyses the input data using the activation function, which is the functions performed by the neuron. This leads to the output data (Fig. 3A). The nodes are connected in series or in parallel to form a network. This network contains one input layer, several hidden layers, and one output layer (Fig. 3B) [61]. Over recent years, deep learning methods like CNNs [62] and recurrent neural networks (RNNs) [63] have been applied into the relation extraction field and have led to promising results.

Artificial Neural Networks (ANNs) are comprised of numerous interconnected computational neurons. These work in an entwined manner to distribute data analysis tasks. This allows ANNs to collectively act together in order to analyse data. Initially the ANN uses the provided data to learn and optimize the analysis process. The basic structure of an ANN can be modelled as shown in Fig. 3B [64]. There are many variables involved in the ANN approach, these include the input, which is normally in the form of a multidimensional vector. This input is distributed to the hidden layers, where decisions are made based on the initial analysis performed in the input layer it then decides based on learned skills whether any of these changes made to the data worsens or improves the final output [65].

Convolutional neural networks (CNNs) are a type of ANN that consists of self-optimizing neurons that accomplish this optimization through learning. Additionally, like ANNs each neuron receives input and performs one or multiple operations. However, they differ from ANNs as they are more commonly used in pattern recognition in images. CNNs are comprised of neurons organised into three layers, the convolutional, pooling and connected layers. CNNs also function in three dimensions-this is due to these networks being focused on image analysis, so the input contains information for height, width, and depth. Each layer of the network contains smaller numbers of neurons that were in the previous layer, however, all of the neurons in preceding layer are connected to the neurons in the next layer. The neurons in the next layer are only connected to a small fraction of neurons within the previous layer [64]. Networks such as CNN which have multiple layers are classed as Deep Neural Networks (DNNs). DNNS are commonly used to analyse transcriptional response datasets. Since these networks consist of multiple layers including many hidden layers, it allows them to be flexible (reviewed in Ref. [66]). DNNs have been used to assist in the creation of a causal gene prediction method based on multiple omics data for ovarian cancer. This model proved to be accurate and effective [67]. Recurrent neural network (RNN) are normally used for NLP. These networks treat inputs and outputs as they are dependent on each other by remembering previous functions, and the results of previous inputs which it then applies to the data analysis. For instance, in language prediction RNNs would try and predict the nest word based on the previous words learned from past experiences [68].

3.2. Learning models and decision tools

There are multiple learning models and methods that can be used to teach an AI. These include supervised, unsupervised, semi-supervised, hybrid and kernel-based learning approaches [69–73]. Supervised approaches require training of AI using labelled training data while unsupervised learning approaches learn by the inherent structure within the data and include methods such as association mining [74]. A method that relies on discovering correlations, or causal structures among sets of attributes in data sets. Semi-supervised learning uses weakly labelled data and includes methods such as the distant supervision approach [75]. Distant supervision relies on the application of a set of rules that

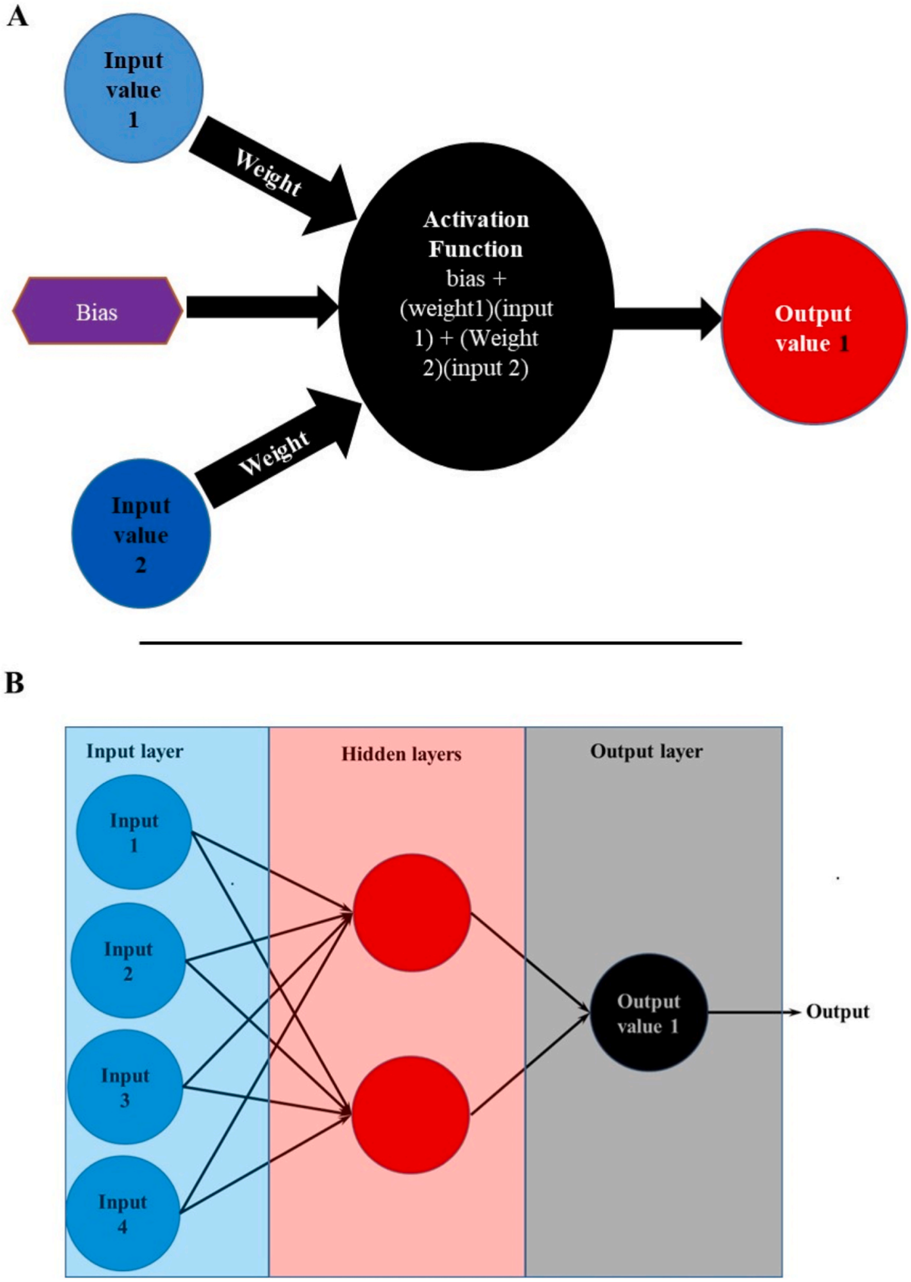

Fig. 3. Neural networks. (A) A single node of a neural network showing the input data has a different weight, which introduces a bias. The activation function is the algorithm acting on the input data and is followed by the generation of output data. (B) Is a representation of an Artificial Neural Network showing the different layers present in the network.

cover the association between different items. Hybrid learning integrates multiple approaches for an AI to learn from data [76].

Decision tools are used in AI to make a decision based on the data analysed. Additionally, they are used as learning models. Decision trees are flowchart like representations of a decision-making process. Each node represents a test performed on the data and each branch is a result of that test, with the terminal node normally represents a classification or label. Tools such as decision trees, random forest and neighbour joining tools are all methods of decision making. A test is then performed on this decision, resulting in another outcome branch. This continues until an end node is reached (Fig. 4 A). This end node is user-defined. Alternately, some nodes are chance nodes, where decisions cannot be made. The decision on where to split a node is normally made using algorithms and results in similar data being grouped together (Fig. 4B) [77]. Random decision forests create multiple decision trees at the same time as it is learning from the data. The output data is the result, or decision, produced by most of the trees. These trees are commonly described as black boxes due to their ability to generate predictions with little user input (Fig. 4C). Therefore, it is not always clear what the algorithm has done to the data or how it reached its conclusions [78].

In a regression analysis one variable is denoted as the independent variable and another as the dependent variable. Additionally, the dependent variable is continuous and terminal nodes represent the mean of the results of the preceding nodes and branch [77]. Neighbour joining is a method used to construct trees where those outputs that are most similar are grouped together as a way of organizing the decision tree into an organized structure. While not a decision tree or method, this looks very similar to the neighbour joining method of constructing phylogenetic trees. This method is a bottom up method as these multiple input nodes then work backwards and are linked by branches to a node most

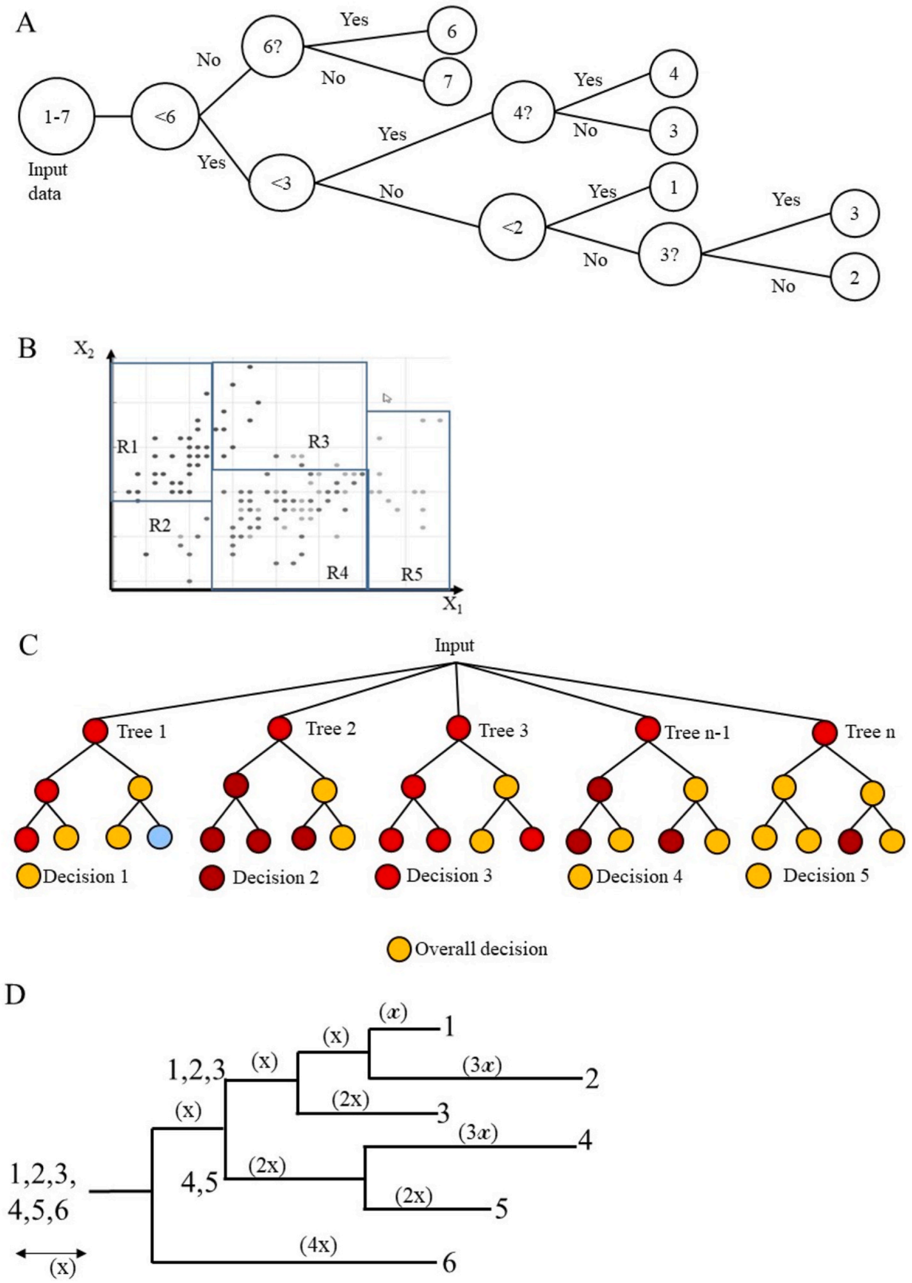

Fig. 4. Schematics representing different decision trees. (A) A basic decision tree (B) Representation of branch splitting techniques (C) Random Forest decision tree (D) a Neighbour joining tree. In (A) each node represents as specific test while the branch splits the data based on the result of the test. (B) demonstrates how the decisions are made to split the data based on how the data is associated. The random forest test (C) increases the accuracy of the decision process by producing many trees and selecting the most common outcomes as the final decision. The neighbour joining tree uses a different method to group the outcomes of any analysis performed on the data based on how similar the outcomes are to each other.

similar to the nodes following it (Fig. 4D) [79].

4. Modelling the effect of genomic mutations on protein structure

All the changes in the genome and transcriptome that occur in cancers, result in changes in the expression of proteins. The proteome is the endpoint of all gene expression, and the proteome can be said to give rise to the observable phenotype of a patient. The proteome of an individual changes as gene expression changes [80]. Changes in gene expression, and mutations in the protein coding regions of genes can change the proteome and therefore the phenotype of cancer cells. These mutations can result in changes in the amino acid sequence resulting in changes in the shape and function of proteins. The effect of these functional changes on the shape of the protein can be predicted by AI using ML and DL to interpret genomic sequence data. Some of the *in silico* modelling programs used to achieve this include SIFT, PANTHER-PSEP and PolyPhen2 [81]. Neural networks are trained to model the changes caused by amino acid changes resulting from genomic mutations. To do this, the network must be able to calculate the changes in secondary structure caused by these amino acid changes and calculate the distances between pairs of residues. The neural network then arranges these secondary structural elements into a three-dimensional structure using specific algorithms, e. g., the Monte Carlo Metropolis algorithm [82]. One of the major barriers to the creation of protein models is the positioning of unstructured loops between secondary structural elements. This process is improved using automated modelling by determining the positions of all non-hydrogen atoms within the loop. The energy of the interactions created by the position of these atoms is estimated through the energy constraints arising from spatial restraints, bond length, bond angle, and improper dihedral angle terms [83]. The final stage of protein modelling is the use of molecular dynamic modelling where the algorithms determine the possible protein conformational changes and this allows the precise shape of the protein (Fig. 5) [84]. One of the DNNs used to assist protein

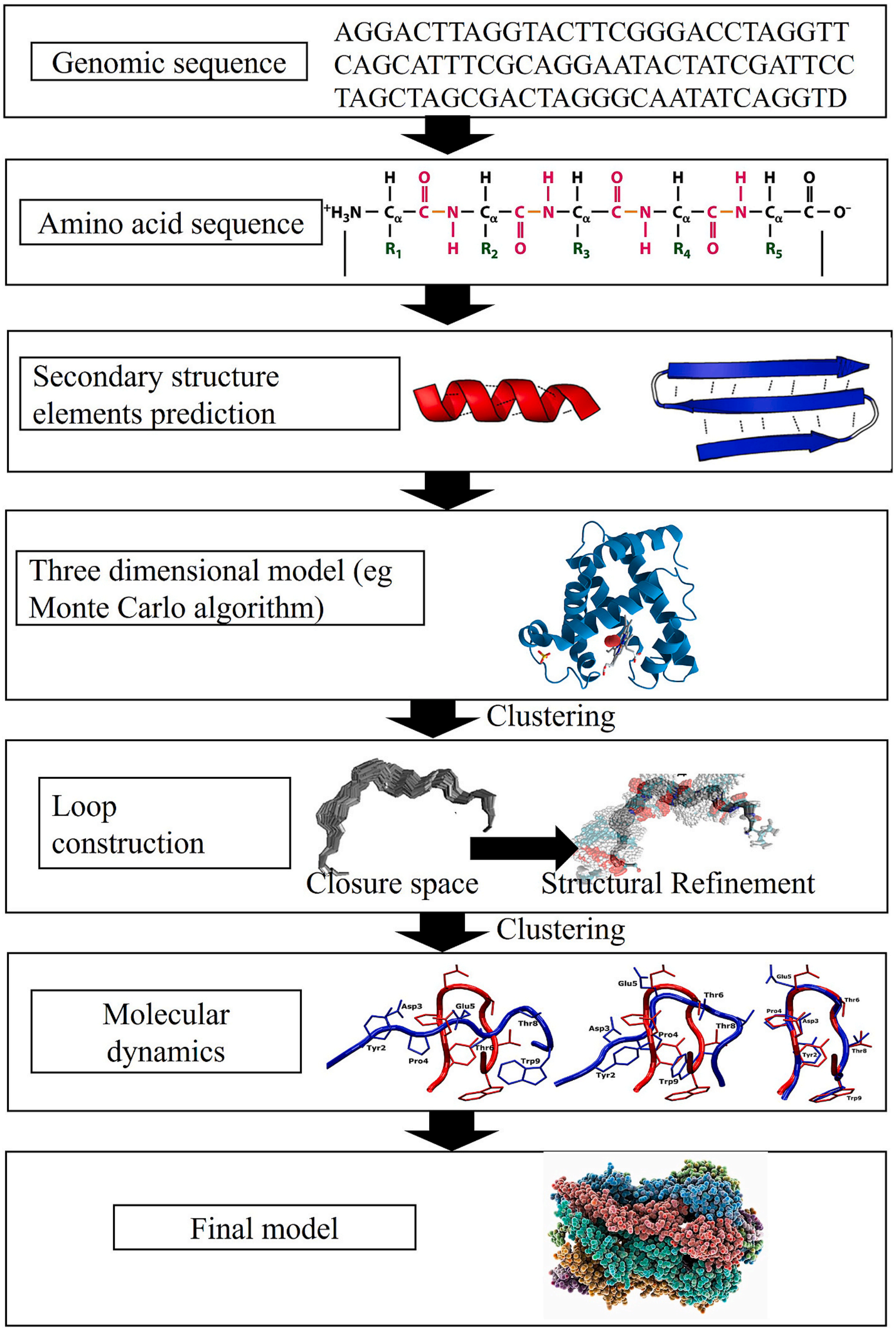
Genomic sequence
AGGACTTAGGTACTTCGGGACCTAGGTT
CAGCATTTCGCAGGAATACTATCGATTCC
TAGCTAGCGACTAGGGCAATATCAGGTD
Amino acid sequence
Secondary structure elements prediction
Three dimensional model (eg Monte Carlo algorithm)
Clustering
Loop construction
Closure space
Structural Refinement
Clustering
Molecular dynamics
Tyr2
Asp3
Pro4
Glu5
Thr6
Thr8
Trp9
Final model

Fig. 5. Typical protein structure prediction pipeline. The gene sequence is translated into an amino acid sequence. Machine learning algorithms are used to predict the secondary structure elements (SSEs) of the protein. These are then arranged in the three-dimensional space using different algorithms. Loop construction is then performed using special algorithms. Molecular dynamics simulations lead to the generation of a three-dimensional protein model (tertiary structure) with predictions on how this structure can change shape during interactions (quaternary structure).

modelling is known as AlphaFold which functions by analysing neighboring amino acid distances and the angles of the peptide bonds between these amino acids [85].

However, the variability in the models that can be generated in this method is large with many slightly to quite different models being generated by the same AI tool. The quality of these models is assessed using assessment programs (MQAPs). Some of these analysis tools are shown in Table 1. Consensus quality assessment involves ranking models in a variety of ways. For instance, the 3D-Jury software examines the generated models and groups them pairwise based on how similar they are. All pairs are assigned a score based on how similar they are. These scores are added together for all pairs generated, and the model with the highest score is accepted as the final model as it is the most consistently similar to other models [86]. Another approach used by software such as QMEANclust, is to use reference models that are selected based on their quality score – QMEAN. The models generated are then compared to the reference models and assigned a global quality score based on the average of its similarity score [87]. Single model quality assessment produces a quality score based on only the single model without comparison to other models. Some of these models are purely statistical analyses while others make use of physical parameters [88]. The PROVE analysis tool uses atom and residue volumes from protein data bank (PDB) structures. Three dimensional structures are inferred from deviations in these traits compared to the expected or normal values. Interacting residues are also identified and placed into one of five separate classes The likelihood of these different classes being correct is established via statistical analysis [89]. Verify3D uses neural networks to predict the probability of finding a specific amino acid in the position predicted by the model [90]. The ProQ assessment tool uses neural networks to assess a model based on multiple physical features of

Table 1
Various consencus and specific model tools for the assesment of protein model quality.

Tool	Basis for assessment	Ref
Consensus Methods		
3D Judge	Pairwise similarity	[86]
Mod-FOLDclust	Pairwise similarity and superimposing of paired models to estimate local accuracy	[86]
Pcons	Pairwise similarity and superimposing of paired models to estimate local accuracy	[92]
QMEANclust	Compare models against a list of reference sequences and assign score based on average similarity score	[87]
Single model methods		
PROVE	Uses atom and residue volumes to infer structure	[89]
ConQuass	Statistical analysis based on evolutionary conservation	[93]
Verify3D	Evaluates model based on solvent accessible area and polarity of residues	[90]
TUNEn	Evaluates structure verus sequence based on local environment	[94]
Verify3D	probability of finding a given amino acid in certain position in a specified environment	[90]
ProQ	Neural network assessment of physical features of the model	[91]
Protein structure function	Structural similarity combined with the structure of proteins that have similar function to vakidate models	[95]

the model. These include the interactions between atoms and residues, solvent accessibility and the secondary structure prediction created in the process of building the model [91].

The *braf* gene encodes the B-RAF protein, a member of the RAF kinase family. Members of this family are responsible for transducing growth signals and is the target of RAF monomer inhibitors and RAF dimer (type II) inhibitors. However, tumor cells resistant to these drugs have shown mutations in the *braf* gene that give rise to structurally altered forms of the protein. An analysis of these mutants was performed using AI to identify mutations that play a role in this drug resistance. It was found that these mutations were found in the kinase domain, and the 3D modelling of the BRAF protein led to the development of new inhibitors [96]. The B-RAF inhibitor PLX4032 (RG7204) was designed using structural information and structural predictions made by AI. The structure-guided discovery of this inhibitor of the B-RAF kinase activity was done using the known 3D structure of B-RAF as determined by x-ray crystallography [97]. Protein modelling algorithms were then used to predict how the specific mutations that give rise to drug resistance, would alter the 3D structure of this protein. These models were then used to design PLX4032 [96].

5. Imaging genomics (Radiomics/Radiogenomics)

The term 'Imaging genomics' also known as radiomics or radiogenomics describes the association of features of a tumor identified through tumor imaging with genomic data such as mutations, copy number variation and gene expression profiles [98]. These features identified in an image include details such as structures, shapes, lines, points, colours, boundaries or even the area of the image closely associated with one of these features. In medical imaging, these features are used to distinguish tumor tissue and normal tissue. The ability to make these distinctions has traditionally only been able to be performed by a human operator meaning it is subject to bias and interpretation and different individuals analyzing the same image commonly obtain different results [99].

However, as AI has advanced it has become able to analyse medical images without the need for human interference. AI and deep learning have led to medical imaging becoming automated and consistent. It is now generally thought that AI can outperform experienced pathologists in the use of medical imaging to diagnose cancer or make prognostic predictions [100]. Imaging genomics relies on the use of AI to extract features identified on an image and link these features with phenotypes. This phenotype reflects protein expression which can then be associated with genomic, transcriptomic and epigenomic changes. This association can then be used to improve diagnostic and prognostic approaches [101]. Therefore, medical imaging can be used to infer that these genetic changes are present within the tumor being imaged. Therefore, these image features can also be used as predictors of survival or indicators of the effects of treatment and even as a more accurate diagnostic tool than conventional medical imaging. As previously stated, CNNs were shown to be especially useful in image analysis and as such have now been specifically applied to medical imaging. CNNs can also perform feature extraction, selection, and classification across different layers [102].

An example of the use of radiogenomics/radiomics is given by the work of Yin et al. (2022). They used an AI brain metastasis detection system. This system used a multi-scale cascaded convolutional network to analyse 3D-enhanced T1-weighted magnetic resonance images. The results produced by this system were compared to those generated by three experienced and three junior radiologists. The system was able to detect brain tumor metastasis with a higher sensitivity and accuracy rates than the six radiologists were [103]. The model was also able to incorporate new data to assist in making predictions on the outcome of treatment. This study therefore demonstrates the valuable role that radiogenomics can play as a diagnostic tool, a treatment monitor and prognostic tool. This highlights the non-invasive nature of radiogenomics, and its low cost, and shows that the field of radiogenomics deserves further investigation [103]. Radiomics/radiogenomics has also been used to stratify the risk of mantle cell lymphoma (MCL) using CT-derived 3D images. Features such as uniformity, entropy, skewness, and difference in entropy were selected and used to detect high risk MCL. Using these features the AI was more reliable in predicting if the MCL patient would have a poor outcome, compared to the use of traditional size measurements of the tumor as a prgnstic tool [104].

Currently, one of the major problems with radiogenomics is the lack of any standard AI system, with each team using a different feature selection process. Despite this, a comparison of various radiogenomic processes, different AI or feature selection, reveals that they come to the same or similar conclusions [105].

6. AI in the diagnosis and treatment of cancer

Cancer is a complex disease and conservative estimates put the number of parameters that need to be considered for correct and accurate medical decision-making at approximately 10 000 parameters [106]. This is obviously too large a number for any physician to fully apply to even a single patient. AI is the obvious solution to this problem, helping to provide faster and more accurate interpretations of patient genomic and transcriptomic data [107].

6.1. Biomarker discovery

One of the most promising techniques for cancer detection is the use of molecular biomarkers. These biomarkers can be used for diagnosis, prognosis and the estimation of patient survival. Biomarkers can also be used to improve cancer treatment and management. Biomarkers can also be used to classify cancers into types or subgroups. Additionally, they can be used to stratify cancers based on the stage of disease. This is important because different types, subgroups, or stages, require different treatments [108]. This involves determining if the presence of these biomarkers is associated with specific cancers, different stages of these cancers or with different patient outcomes. In these cases, relevant biomarkers have been identified through different omics technologies and the analysis of this data using AI [108]. Additionally, drug resistance is a major obstacle for the successful treatment of cancer. The identification of genes, epigenetic changes, and the pathways responsible for the development of drug resistance would assist in solving this problem. Once identified, these changes can be targeted, and new drugs can be developed to counteract these changes. For example, analysis of the genomic data revealed that mutations in the estrogen receptor (ESR1) are responsible for secondary resistance in breast cancer [109].

Physical biopsies require tissue to be removed and examined histologically, whereas the examination of molecular biomarkers via NGS can be non-invasive and performed through blood tests. A good example of a molecular biomarker is the circulating cancer antigen 125, which is used for the detection of ovarian cancer [110]. Liquid biopsies are a way to carry out non-invasive diagnosis for cancers. They can also improve prognostics and assist in drug-response monitoring [111]. This requires the identification and characterization of novel biomarkers. NGS can assist in this process by detecting the presence of mRNAs or miRNAs specific to a particular cancer, but also by detecting mutation signatures and tumor mutational burden (TMB). Advanced statistical and data analysis needs to be applied to all these changes detected by the analysis of NGS data using AI [111].

RNA sequencing provides information on changing gene expression signatures in cancer as well as to detect mutations in RNA which can affect gene expression through changes in the final protein sequence. Both these changes are related to the underlying molecular mechanisms of cancer and both changes can be used as biomarkers for not only the detection of different types of cancer, but also the staging of the tumor. These changes can also act as biomarkers for the patient's prognosis or response to treatment [13].

6.2. Problems facing the application of AI to cancer diagnosis, prognosis and treatment

Multiple studies have demonstrated how AI can outperform humans in the interpretation of the vast quantities of data pertaining to a complex disease such as cancer. However, it is important to remember that AI should always be used to augment human intelligence and not replace it. This means the outcomes of any analysis performed by AI should be assessed by qualified experts in their respective fields. Machine and deep learning by AI must also be assessed or supervised by experts in bioinformatics and programming [112]. One of the biggest problems facing the application of AI and DL to cancer diagnosis, prognosis and treatment is the aforementioned black box problem. Essentially this concerns a lack of knowledge concerning what the AI system is actually doing and how it comes to its final conclusion. Once an AI is fully automated and requires no human intervention, it may become uncertain how an AI is selecting features or making decisions. This may create doubt as to the accuracy of the predictions made and force clinicians and researchers to accept these results on "blind faith" [113]. Researchers have also tried to develop AI systems whose actions can be understood by the physicians and clinicians that are using it. Kweng et al. (2022) developed an AI using ML to predict if prostate cancer patients could effectively be treated using nerve-sparing radical prostatectomy. The AI did this by predicting whether a tumor could extend beyond the prostate. All decisions and analyses produced by the AI could be analysed and explained in layperson terms using a publicly available web application, Shapley Additive exPlanations [114].

Deep learning also requires a large amount of data to learn enough to generate algorithms that it can be applied to new data. In order to get this learning data, cancer research studies require multiple samples to act as training data [115]. Additionally, the use of big data and AI poses ethical problems since it uses the patients data, in some cases to accomplish tasks not directly related to the care of the patient and the use of this data may not always occur with the consent of the patient [116].

7. Conclusions

Precision medicine promises to offer unprecedented levels of patient care and cancer treatment. It would allow not only for precise treatment based on the patient's lifestyle, mutation profile and protein expression patterns but would also provide the most accurate information regarding the patient's ethnicity and family history when it comes to risk factors and treatment response (Fig. 6). Currently a patient's ethnicity is judged on their self-reporting or their appearance and may disregard the actual genetic background of a patient. In order to do this with the highest accuracy a "digital twin" of the patient would be required. This in turn requires that large amounts of data describing the patient's lifestyle and biology be captured and curated. The management and application of the increasingly large sets of data required to make a digital twin for the purpose of assisting in the control of cancer, cannot be performed timeously and accurately without the aid of AI. The ability of AI to assist in monitoring this response of a patient to a specific treatment, monitor the recovery of the patient and finally predict treatment outcomes, means that these treatments can be fine-tuned to suit the situation (Fig. 6).

The problems facing the implementation of AI and the use of big data are not insurmountable. Ethical problems with the use of big data can be solved through policy makers and the implementation of simple rules and guidelines governing its use. The problem of AI being a mysterious black box that clinicians and oncologists would be uncomfortable using or trusting is being solved with the development and implementation of methods to test the accuracy of the predictions made by the system as

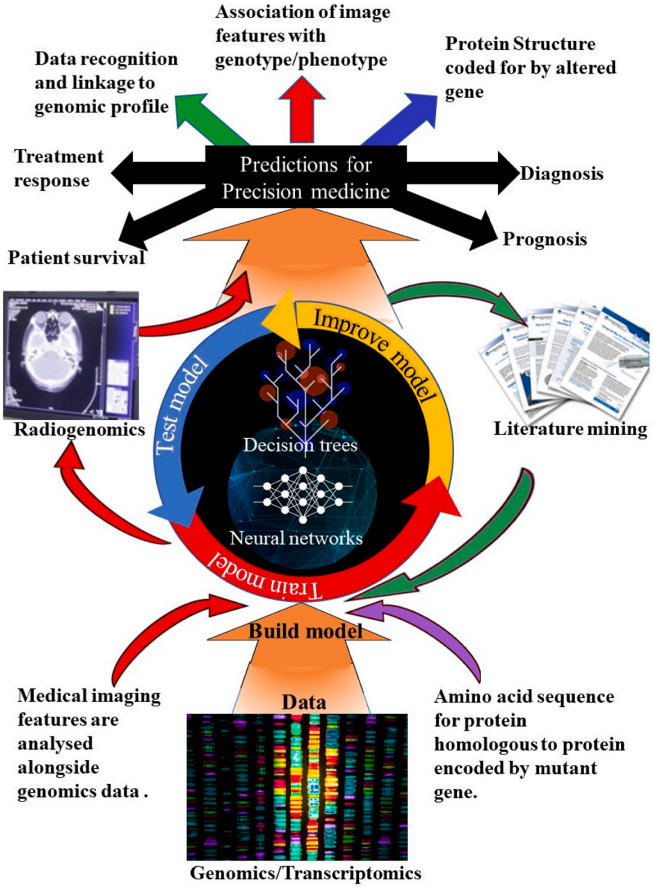
Association of image features with genotype/phenotype
Data recognition and linkage to genomic profile
Protein Structure coded for by altered gene
Treatment response
Predictions for Precision medicine
Diagnosis
Patient survival
Prognosis
Improve model
Test model
Decision trees
Neural networks
Train model
Radiogenomics
Literature mining
Build model
Medical imaging features are analysed alongside genomics data .
Data
Amino acid sequence for protein homologous to protein encoded by mutant gene.
Genomics/Transcriptomics

Fig. 6. The application of AI to precision medicine: Genomic data can be combined with multiple types of associated data such as published literature, medical images, wild type protein models and gene expression data. AI algorithms use neural networks and learning algorithms to create, test and improve models to achieve accurate predictions. The end result will be the association of features or patterns in this data with markers for diagnosis, prognosis, treatment outcome prediction and patient survival. Additionally, the genotype can be associated with specific phenotypic features or used to search literature on specific genomic profiles or genes. Finally, mutations present in the gene sequences can be used to predict changes in protein structure.

well as revealing some of the decision-making processes made by the AI system. The problem of the availability of training data will solve itself as more studies are performed and data collected. This data can be used retrospectively to train the AI and as such this problem will be solved by the passage of time. Lastly the enormous amounts of data storage capability required to keep the genomic, transcriptomic, proteomic, and medical record data for every patient should not represent a problem, as computer hardware development continue to advance. In summary the constant development of AI and computational algorithms promises to provide a far better outlook for cancer patients. Using personalized, clinically obtained genomics data from the patient, analysed by AI, cancer screening and diagnosis can be improved leading to the prevention of serious disease. At the same time analysis of this data by AI systems can result in more targeted treatments and the improved monitoring of treatment results leading to better patient care, which is the aim of precision oncology.

Author contribution

Zodwa Dlamini- Study design, Editing and rewrites, Amanda Skepu-Editing and rewrites, Namkug Kim- Editing and rewrites, Mahlori Mkhabele-writing, Richard Khanyile- Writing, Thulo Molefi-writing Editing and rewrites, Sikhumbuzo Mbatha Editing and rewrites, Botle Setlai- Writing, Thanyani Mulaudzi - Editing and rewrites, Mzubanzi Mabongo Editing and rewrites, Meshack Bida- Editing and rewrites, Minah Kgoebane-Maseko- Editing and rewrites, Kgomotso Mathabe-Editing and rewrites, Zarina Lockhat- - Editing and rewrites, Mahlatse Kgokolo - Editing and rewrites, Nkhensani Chauke-Malinga - Editing and rewrites, Serwalo Ramagaga - Editing and rewrites, Rodney Hul-Writing.

Conflicts of interest

We declare we have no conflict of interest.

Acknowledgement

We would like to thank the South African Medical Research Council (SAMRC) Grant Number 23108 and the National Research Foundation (NRF) Grant Number 138139, for funding this research.

References

[1] Batch KE, et al. Developing a cancer digital twin: supervised metastases detection from consecutive structured radiology reports. Front. Artif. Intelligence 2022;5. 826402-826402.
[2] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;69s: S36–s40.
[3] Aarvik P. Artificial Intelligence–a promising anti-corruption tool in development settings. Available online, https://beta.u4.no/publications/artificial-intelligence-a-promising-anti-corruption-tool-in-development-settings.pdf (accessed on 12 December).
[4] Bersanelli M, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinf 2016;17(Suppl 2):15. Suppl 2.
[5] Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinf 2018;19(1):202.
[6] Szymczak S, et al. Machine learning in genome-wide association studies. Genet Epidemiol 2009;33(Suppl 1):S51–7.
[7] Telenti A, et al. Deep learning of genomic variation and regulatory network data. Hum Mol Genet 2018;27(R1):R63–r71.
[8] Chari R, et al. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. BMC Syst Biol 2010;4:67.
[9] Wang B, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11(3):333–7.
[10] Syrjala KL. Opportunities for improving oncology care. Lancet Oncol 2018;19(4): 449.
[11] Pennell NA, et al. Economic impact of next generation sequencing vs sequential single-gene testing modalities to detect genomic alterations in metastatic non-small cell lung cancer using a decision analytic model. American Society of Clinical Oncology; 2018.
[12] Punetha J, Hoffman EP. Short read (next-generation) sequencing: a tutorial with cardiomyopathy diagnostics as an exemplar. Circ Cardiovasc Genet 2013;6(4): 427–34.
[13] Wang Y, et al. Changing technologies of RNA sequencing and their applications in clinical oncology. Front Oncol 2020;10:447.
[14] Lyu B, Haque A. Deep learning based tumor type classification using gene expression data. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics; 2018.
[15] Bartsch Jr G, et al. Use of artificial intelligence and machine learning algorithms with gene expression profiling to predict recurrent nonmuscle invasive urothelial carcinoma of the bladder. J Urol 2016;195(2):493–8.
[16] Pepke S, Ver Steeg G. Comprehensive discovery of subsample gene expression components by information explanation: therapeutic implications in cancer. BMC Med Genom 2017;10(1):12.
[17] Li MM, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the association for molecular pathology, American society of clinical oncology, and college of American pathologists. J Mol Diagn 2017;19(1):4–23.
[18] Gargis AS, et al. Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nat Biotechnol 2015;33(7):689–93.
[19] Pezoulas VC, et al. Machine learning approaches on high throughput NGS data to unveil mechanisms of function in biology and disease. CANCER GENOMICS PROTEOMICS 2021;18(5):605–26.
[20] Meng C, et al. Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings Bioinf 2016;17(4):628–41.
[21] Raimundo F, Vallot C, Vert JP. Tuning parameters of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol 2020;21(1):212.
[22] Morin O, et al. An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. Nat Can (Que) 2021;2(7):709–22.
[23] Bewicke-Copley F, et al. Applications and analysis of targeted genomic sequencing in cancer studies. Comput Struct Biotechnol J 2019;17:1348–59.
[24] Kothen-Hill ST, et al. Deep learning mutation prediction enables early stage lung cancer detection in liquid biopsy. 2018.
[25] Wood DE, et al. A machine learning approach for somatic mutation discovery. Sci Transl Med 2018;10(457).
[26] Bathke J, Lühken G. OVarFlow: a resource optimized GATK 4 based Open source Variant calling workFlow. BMC Bioinf 2021;22(1):402.
[27] Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7(4):248–9.
[28] Vaser R, et al. SIFT missense predictions for genomes. Nat Protoc 2016;11(1):1–9.
[29] Calabrese R, et al. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 2009;30(8):1237–44.
[30] Porta-Pardo E, et al. Comparison of algorithms for the detection of cancer drivers at subgene resolution. Nat Methods 2017;14(8):782–8.
[31] Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. Bioinformatics 2008;24(20):2397–8.
[32] Ferrer-Costa C, et al. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 2005;21(14):3176–8.
[33] Qi H, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun 2021;12(1):510.
[34] Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 2015;31(5):761–3.
[35] Capriotti E, et al. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. Hum Mutat 2008;29(1):198–204.
[36] Karchin R, et al., LS-SNP. large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 2005;21(12):2814–20.
[37] Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol 2006;356(5):1263–74.
[38] Dobson RJ, et al. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinf 2006;7:217.
[39] Krishnan VG, Westhead DR. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 2003;19(17):2199–209.
[40] Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 2005;21(10):2185–90.
[41] Carter H, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 2009;69(16): 6660–7.
[42] Kaminker JS, et al. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res 2007;35:W595–8 (Web Server issue).
[43] Andreoletti G, et al. Reports from the fifth edition of CAGI: the critical assessment of genome interpretation. Hum Mutat 2019;40(9):1197–201.
[44] Zomnir MG, et al. Artificial intelligence approach for variant reporting. JCO Clin Cancer Inform 2018;2.
[45] Corti G, et al. A genomic analysis workflow for colorectal cancer precision oncology. Clin Colorectal Cancer 2019;18(2):91–101.e3.
[46] Levsky JM, et al. Single-cell gene expression profiling. Science 2002;297(5582): 836–40.
[47] Jain M, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 2018;36(4):338–45.
[48] Adamopoulos PG, et al. Targeted long-read sequencing decodes the transcriptional Atlas of the founding RAS gene family members. Int J Mol Sci 2021;22(24):13298.
[49] Soomro PD, et al. Bio-NER: biomedical named entity recognition using rulebased and statistical learners8; 2017. p. 163–70.

[50] Cheng D, et al. Poly Search: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res 2008;36:W399–405 (Web Server issue).

[51] de Ridder J, et al. Co-occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes. Bioinformatics 2007;23(13):i133–41.

[52] Hakenberg J, et al. A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions. J Biomed Inf 2012;45(5):842–50.

[53] Kehl KL, et al. Artificial intelligence-aided clinical annotation of a large multi-cancer genomic dataset. Nat Commun 2021;12(1):7304.

[54] Samuel ALJI, development. Some studies in machine learning using the game of checkers. J.o.r 1959;3(3):210–29.

[55] Uddin S, et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inf Decis Making 2019;19(1):281.

[56] Falk T, et al. U-Net: deep learning for cell counting, detection, and morphometry16; 2019. p. 67–70. 1.

[57] Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: A survey. J. Artif. Intell. Res. 1996;4:237–85.

[58] Banf M, Rhee SY. Computational inference of gene regulatory networks: approaches, limitations and opportunities. Biochim Biophys Acta Gene Regul Mech 2017;1860(1):41–52.

[59] Maleki F, et al. Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment. Neuroimaging Clin 2020;30(4):433–45.

[60] Joshi AV. Machine learning and artificial intelligence. Springer; 2020.

[61] Kuwahara T, et al. Current status of artificial intelligence analysis for endoscopic ultrasonography. Dig Endosc 2021;33(2):298–305.

[62] Lee K, et al. Deep learning of mutation-gene-drug relations from the literature. BMC Bioinf 2018;19(1):21.

[63] Peng B, et al. Recurrent neural networks with external memory for spoken language understanding. In: natural language processing and Chinese computing. Springer; 2015. p. 25–35.

[64] O’Shea K, Nash RJapa. An introduction to convolutional neural networks. 2015.

[65] Baskin II, Winkler D, Tetko IV. A renaissance of neural networks in drug discovery. Expet Opin Drug Discov 2016;11(8):785–95.

[66] AlQuraishi M, Sorger PK. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. Nat Methods 2021;18(10):1169–80.

[67] Ye L, et al. An ovarian cancer susceptible gene prediction method based on deep learning methods. Front Cell Dev Biol 2021;9:730475.

[68] Dupond S. A thorough review on the current advance of neural network structures. Annual Reviews in Control. 2019;14:200–30.

[69] Kim S, et al. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. J Biomed Inf 2015;55:23–30.

[70] Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inf Assoc 2011;18(5):594–600.

[71] Xu Y, et al. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. J Am Med Inf Assoc 2012;19(5):824–32.

[72] Tikk D, et al. A detailed error analysis of 13 kernel methods for protein-protein interaction extraction. BMC Bioinf 2013;14:12.

[73] Yang Z, et al. Multiple kernel learning in protein-protein interaction extraction from biomedical literature. Artif Intell Med 2011;51(3):163–73.

[74] Alicante A, et al. Unsupervised entity and relation extraction from clinical records in Italian. Comput Biol Med 2016;72:263–75.

[75] Quirk C, Poon HJapa. Distant supervision for relation extraction beyond the sentence boundary. 2016.

[76] Muzaffar AW, Azam F, Qamar U. A relation extraction framework for biomedical text using hybrid feature set. Comput Math Methods Med 2015;2015:910423.

[77] Kamiński B, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. Cent Eur J Oper Res 2018;26(1):135–59.

[78] Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. IEEE; 1995.

[79] Saitou N. M.J.M.b. Nei, and evolution, The neighbor-joining method: a new method for reconstructing phylogenetic trees4; 1987. p. 406–25. 4.

[80] Neagu A-N, et al. Proteomics and its applications in breast cancer. Am J Cancer Res 2021;11(9):4006–49.

[81] Tang H, Thomas PD. Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics 2016;203(2):635–47.

[82] Fischer AW, et al. CASP11–An evaluation of a modular BCL::Fold-Based protein structure prediction pipeline. PLoS One 2016;11(4):e0152517.

[83] Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9(9):1753–73.

[84] Karplus M, Kuriyan J. Molecular dynamics and protein function. Proceedings of the National Academy of Sciences 2005;102:6679–85.

[85] Skolnick J, et al. AlphaFold 2: why it works and its implications for understanding the relationships of protein sequence, structure, and function. J Chem Inf Model 2021.

[86] Ginalski K, et al. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19(8):1015–8.

[87] Benkert P, Schwede T, Tosatto SC. QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. BMC Struct Biol 2009;9:35.

[88] Melo F, Feytmans E. Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 1998;277(5):1141–52.

[89] Pontius J, Richelle J, Wodak SJ. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 1996;264(1):121–36.

[90] Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253(5016):164–70.

[91] Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003;12(5):1073–86.

[92] Larsson P, et al. Assessment of global and local model quality in CASP8 using Pcons and ProQ. Proteins 2009;77(Suppl 9):167–72.

[93] Kalman M, Ben-Tal N. Quality assessment of protein model-structures using evolutionary conservation. Bioinformatics 2010;26(10):1299–307.

[94] Lin K, May AC, Taylor WR. Threading using neural nEtwork (TUNE): the measure of protein sequence-structure compatibility. Bioinformatics 2002;18(10):1350–7.

[95] Wilson D, et al. The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res 2007;35:D308–13 (Database issue).

[96] Bollag G, et al. Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. Nature 2010;467(7315):596–9.

[97] Zhang C, et al. RAF inhibitors that evade paradoxical MAPK pathway activation. Nature 2015;526(7574):583–6.

[98] Bodalal Z, et al. Radiogenomics: bridging imaging and genomics. Abdom Radiol (NY) 2019;44(6):1960–84.

[99] Bi WL, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA A Cancer J Clin 2019;69(2):127–57.

[100] Gürsoy Çoruh A, et al. A comparison of the fusion model of deep learning neural networks with human observation for lung nodule detection and classification. Br J Radiol 2021;94(1123):20210222.

[101] Rutman AM, Kuo MD. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. Eur J Radiol 2009;70(2):232–41.

[102] Berger MF, Mardis ER. The emerging clinical relevance of genomics in cancer medicine. Nat Rev Clin Oncol 2018;15(6):353–65.

[103] Yin S, et al. Development and validation of a deep-learning model for detecting brain metastases on 3D post-contrast MRI: a multi-center multi-reader evaluation study. Neuro Oncol; 2022.

[104] Lisson CS, et al. Longitudinal CT imaging to explore the predictive power of 3D radiomic tumour heterogeneity in precise imaging of mantle cell lymphoma (MCL). Cancers 2022;14(2).

[105] Wang Y, et al. Radiomics and radiogenomics in evaluation of colorectal cancer liver metastasis. Front Oncol 2021;11:689509.

[106] Abernethy AP, et al. Rapid-learning system for cancer care. J Clin Oncol 2010;28(27):4268–74.

[107] Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. NPJ Digit Med 2018;1:5.

[108] Dlamini Z, et al. Artificial intelligence (AI) and big data in cancer and precision oncology. Comput Struct Biotechnol J 2020;18:2300–11.

[109] Kumar S, et al. Tracking plasma DNA mutation dynamics in estrogen receptor positive metastatic breast cancer with dPCR-SEQ. NPJ Breast Cancer 2018;4:39.

[110] Wang J, et al. RNA sequencing (RNA-Seq) and its application in ovarian cancer152; 2019. p. 194–201. 1.

[111] Palmirotta R, et al. Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology. Ther Adv Med Oncol 2018;10:1758835918794630.

[112] Pashkov VM, Harkusha AO, Harkusha YO. Artificial intelligence in medical practice: regulative issues and perspectives. Wiad Lek 2020;73(12 cz 2):2722–7.

[113] Sorell T, Rajpoot N, Verrill C. Ethical issues in computational pathology. J Med Ethics 2021.

[114] Kwong JCC, et al. Explainable artificial intelligence to predict the risk of side-specific extraprostatic extension in pre-prostatectomy patients. Can Urol Assoc J; 2022.

[115] Hussain Z, et al. Differential data Augmentation techniques for medical imaging classification tasks. AMIA Annu Symp Proc 2017;2017:979–84.

[116] Rigby MJJ. Ethical dimensions of using artificial intelligence in health care. A.J.o.E. 2019;21(2):121–4.