

Review

An overview of artificial intelligence in the field of genomics

Khizra Maqsood¹ · Hani Hagraš¹ · Nicolae Radu Zabet²

Received: 17 November 2023 / Accepted: 2 January 2024

Published online: 30 January 2024

© The Author(s) 2024 [OPEN](#)

Abstract

Artificial intelligence (AI) is revolutionizing many real-world applications in various domains. In the field of genomics, multiple traditional machine-learning approaches have been used to understand the dynamics of genetic data. These approaches provided acceptable predictions; however, these approaches are based on opaque-box AI algorithms which are not able to provide the needed transparency to the community. Recently, the field of explainable artificial intelligence has emerged to overcome the interpretation problem of opaque box models by aiming to provide complete transparency of the model and its prediction to the users especially in sensitive areas such as healthcare, finance, or security. This paper highlights the need for eXplainable Artificial Intelligence (XAI) in the field of genomics and how the understanding of genomic regions, specifically the non-coding regulatory region of genomes (i.e., enhancers), can help uncover underlying molecular principles of disease states, in particular cancer in humans.

1 Introduction

In 1957, Francis Cricks proposed the central dogma of molecular biology which explains the flow of genetic information in a living organism summarized in a pathway (Fig. 1) from DNA (Deoxyribonucleic Acid) to RNA (Ribonucleic Acid) and from RNA to protein (a functional form of the DNA) [12]. DNA has double-helical strands containing four basic units called nucleotides: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). The two strands of DNA are linked with a chemical bond between bases; A is paired with T and C with G. The DNA base sequence contains all biological information to be transcribed into a protein product. In living cells, DNA is organized in the form of chromosomes, which are further organized into segments of DNA called genes that encode for proteins (see Fig. 2). The sum of all genes or genetic material that an organism possesses is known as the genome [15]. The field of life science that focuses on studying the genome or genomic sequences of organisms is called genomics. The human genome possesses approximately 3 billion DNA base pairs, and the field of human genomics aims to link the genome with molecular and physical characteristics [2]. It is a data-driven science that involves high-throughput next-generation sequencing (NGS) technology development that generates data on the whole genome of an organism. These sequencing techniques include whole exome sequencing (WES), whole genome sequencing (WGS), as well as transcriptomic, chromatin, and epigenetic profiling.

In 2001, the completion of the Human Genome Project (HGP) was an important scientific development in the field of genomics, by providing the reference of most of the human genome. Recent technological advances in long reads have allowed for improved human genome reference by sequencing the remaining 8% of the human genome [30]. The sequencing of the whole genome allowed a better understanding of the genetic variation among the organisms

✉ Hani Hagraš, hani@essex.ac.uk; ✉ Nicolae Radu Zabet, r.zabet@qmul.ac.uk | ¹The Computational Intelligence Centre, School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK. ²Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK.



Fig. 1 The Central Dogma of Molecular Biology: Genetic information is transformed from DNA to RNA in the process of transcription. RNA is then translated into the final protein product, which have a variety of functions

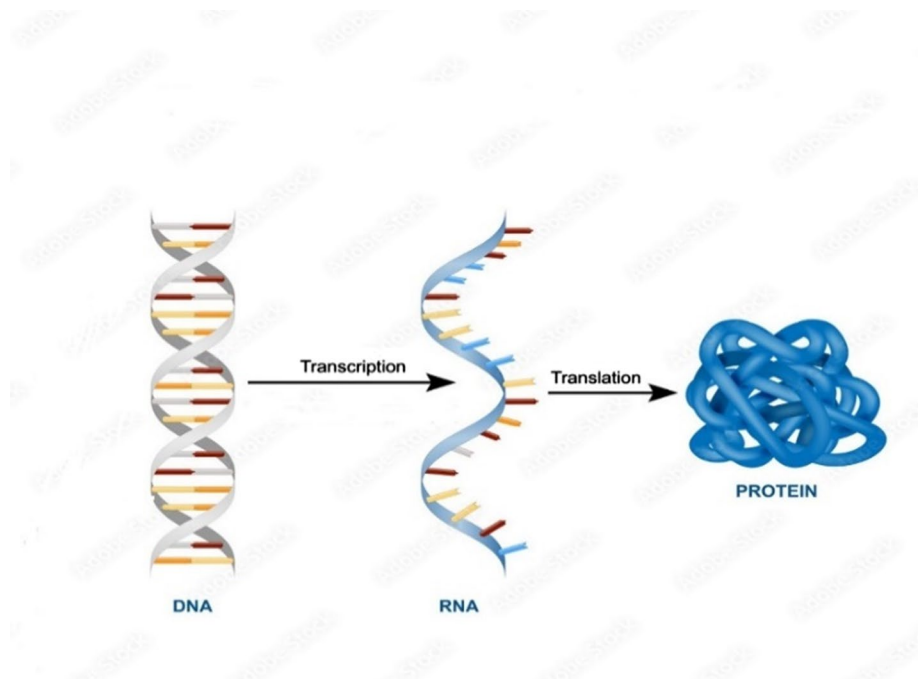
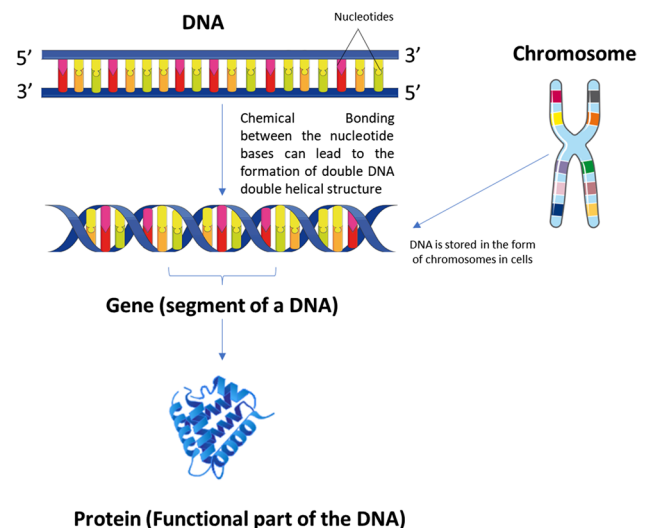
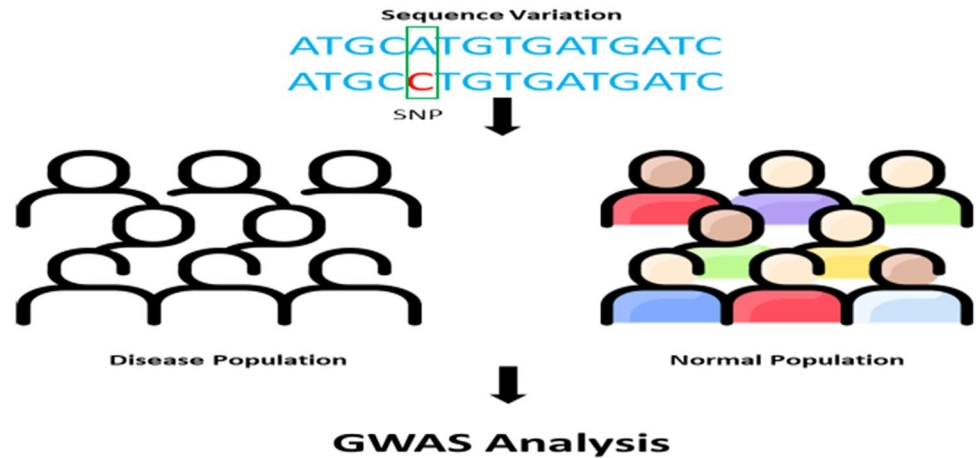


Fig. 2 The Nucleotide bases make a chemical bond to form a double helical structure called DNA. Genetic material is made up of DNA that is tightly packed into chromosomes. Only a certain region of DNA contains genes that code for proteins



or even within the different cells, tissues, and disease states of an organism. The findings of the HGP suggested that all humans are 99.9% genetically identical and only 0.01% variation in the human genome can make all humans phenotypically different, such as their disease susceptibility, responses towards drugs, and physical traits (hair colour, eye colour, height, intelligence, etc.) [14]. One major aim of genomics is to identify the underlying changes or mutations that may occur in DNA sequences to alter cellular processes and cause the disease states, usually done through genome-wide association studies (GWAS) (see Fig. 3). It is worth noting that not all occurring mutations are disease-causing; for example: not all single nucleotide polymorphisms (SNPs) (a single change in a base pair) or indels (insertions or deletions of small pieces of DNA) change the DNA sequence coding for a protein (synonymous mutations) or expression of the genes [18]. If we can identify which variation is linked to a specific disease, we will be able to design better treatments, drugs, or even cures. McGuire *et al.* [26] reported that investigating genetic variation could improve our understanding of why certain people respond differently to the same medications. That is where the personalized medicine concept comes, where pharmacogenomics can develop and prescribe personalized medicine to an individual though understanding their genetic makeup.

Fig. 3 A mutation has been detected in the gene sequence that is responsible for the disease state. For complete analysis, a genome-wide association study is required



Cancer is one of the most prevalent chronic diseases that is caused by genome alteration. Including base substitutions, deletions, rearrangements, or amplifications. The mechanisms of sequence alteration vary between different cancer types.

Furthermore, genomics also plays an important role in managing and understanding infectious diseases on both population and individual levels [24]. In particular, it helps researchers identify and keep track of the emergence of drug resistance in pathogenic organisms. For example, during the COVID-19 pandemic, genomics helped scientists track virus (pathogens) transmission to understand how the strain was evolving to aid the development of effective vaccines. Genomics is also enabling more targeted tests such as for rare disorders, tumour genome analysis, and non-invasive prenatal screening. Furthermore, genomics has also revolutionized the field of agriculture by helping scientists understand the genetic makeup of livestock and crops. Genomics will allow scientists to develop genetically modified organisms (GMO) that can be pest resistant, tolerate harsh environmental conditions and increase yield. This is needed to handle the challenges associated with the growing world population to ensure food security [32]. Biodiversity can also be understood by comparing the genomes of various species and looking at the underlying principle of the evolutionary history of organisms and their adaptation to different environmental conditions.

2 Cancer genomics

Cancer develops because of alterations or mutations that occur in the DNA sequence of genes that regulate cell survival, division, or other hallmarks of the transformed phenotype, resulting in the development of uncontrollable cell growth and the spread of abnormal cells. These cells can acquire genetic mutations that affect normal cell growth mechanisms and lead to the formation of tumours. For several years, researchers have been trying to understand the biological basis of various cancers showing variable clinical outcomes. Genomics can provide insights into the underlying principles of this heterogeneous and complex disease. Genomic studies help researchers identify multiple mutated genes that cause cancer and these are called oncogenes. A common example is the *TP53* gene, which is mutated in different cancers [42]. The use of NGS technologies can provide us with the whole genome profiling of a cancer patient, which can help in identifying and understanding clinically relevant genetic variations that can be targeted for potential therapies. In 1998, the first molecularly targeted drug was introduced based on comprehensive genomic profiling (CGP) called trastuzumab in patients with *ERBB2*-overexpressed breast cancer. Since then, several novel targeted therapies have been discovered, including BRAF melanoma inhibitors, BCR/ABL chronic myeloid leukaemia inhibitors, and epidermal growth factor receptor (EGFR) non-small cell lung cancer tyrosine kinase inhibitors, which provided robust therapeutic responses [29].

Most large-scale genome projects such as the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) have mainly focused on cancer genome characterization. Through genetic variations that occur in the coding part of the genome, they have identified numerous novel mutations. However, only 2% of the human genome is coding, with the remaining 98% being non-coding, and there is very limited information on how variations in the non-coding part of the genome can affect the development of cancer [42]. Elliott & Larsson [16] reported that mutations in the non-coding part of the genome are abundant, but their effects are so far poorly understood. Recent studies have shown that variations in the non-coding regulatory region of the human genome are highly associated with disease conditions. The identification and mechanisms of gene regulatory regions are not only important to understand the

function of the genome, but they also help in the widening the understanding of disease causation and provide a better overview of disease state. For example, a mutation in the regulatory region of the *RB1* gene has been found to be a major source of brain cancer glioblastoma [4]. Genes have multiple non-coding regulatory regions that include promoters, activators, enhancers, and silencers. This paper will focus on enhancer regions, and why identifying and understanding the mechanisms of enhancers is important in cancer.

3 Enhancers and cancer

Enhancers are non-regulatory elements responsible for controlling the transcription of one or more genes (Fig. 4). The characterisation and identification of these regions is important in the context of human disease to determine disease causation and aid in the development of novel drugs. Enhancers can activate expression of proximal or distal genes, the latter function is important in chromosome looping and Topologically Associated Domains [10]. Enhancers are cell-type specific and are located in different regions of the genome [6]. Recent studies have shown that enhancers are associated with several epigenetic markers, including histone modification signals such as H3K4me1/2/3 [11], H3K27ac [13], and H3K9ac [21], cofactors (e.g. cohesion and mediator complex) and, chromatin-modifying molecules (e.g. p300). This histone modification based data can provide the significant evidence in predicting active enhancers. In general there are two types of enhancers: (1) Signal-dependent or inducible enhancers and (2) cell type-specific enhancers. The latter cover the majority of all enhancers present in the genome.

As all human cell types can possess the same genome, cell-type specific enhancers are an important factor to determine cell-type specific gene expression programming. The mammalian genome contains millions of enhancers but only a small number of enhancers are active in each cell type and the activity of enhancers is specific to its targeted gene [22]. The term super-enhancers (Fig. 5) used to represent the active enhancer clusters that are present in a high abundance in a specific genomic region [48]. These enhancers mainly regulate genes that are important for determining the cell identity. Therefore, enhancers provide the basis of cell identity and mutations in these enhancer regions can cause abnormal cell growth and cause several diseases. The fluctuations in DNA methylation are also a cause of cancer development and can directly affect the activity of the enhancers [40]. Furthermore, mutations in regulatory regions of oncogenes can show a major impact in causing brain tumours [4]. In glioblastoma, EGFR amplification is linked with the remodelled enhancer landscapes through the synthesis of FOXG1 and SOX9-dependent transcriptional factors. This signature is highly sensitive to small molecules that disrupt H3K27ac inhibitors. H3K27ac is a histone modification that plays a key role in the epigenetic regulation that controls the gene transcription, enhancer activity and chromatin structure. EGFR amplification in glioblastoma is highly sensitive to small molecules that disrupt H3K27ac inhibitors and

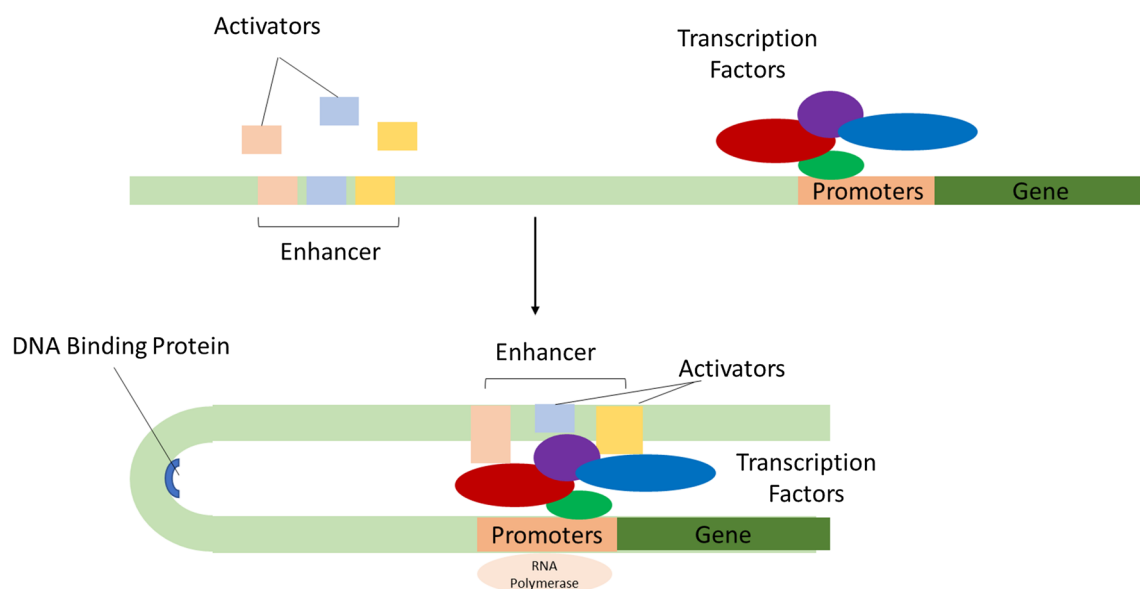


Fig. 4 Enhancers are the non-regulatory elements of DNA. They can make 3D contact with other non-regulatory elements of DNA and are bound by transcription factor proteins to control gene expression [31]

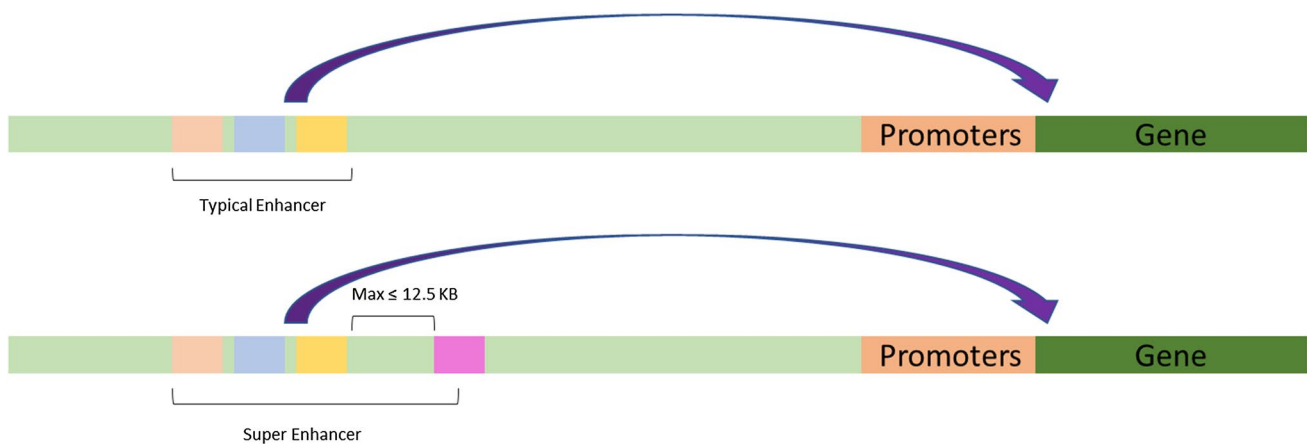


Fig. 5 Systematic representation of typical and super-enhancers

activate an oncogenic gene expression program. This results in the activation of repetitive element expression including an endogenous retroviral element [37]. Furthermore, SMARCB1 is the core subunit of SWI/SNF chromatin remodelling that is lost in cancer which is responsible for maintaining the SWI/SNF complexes and it may result in the disruption of the enhancer-mediated region of genes necessary for cell differentiation [].

Thandapani [43] reported that H3K4me1 histone modification of enhancers is catalyzed by MLL3/MLL4. In various cancer types, MLL3 and MLL4 are mutated, which can reduce the amount of H3K4me1 on enhancers and prevent binding of the mediator complex to those enhancers. MLL4 loss impaired the super-enhancer in lung cancers for the tumour suppressor PER2 gene. Mutation in MLL3 and MLL4 can also lead to therapeutic resistance and dysregulation of enhancers in various cancers. Therefore, a deep understanding of enhancer patterns can help reveal novel activation mechanisms oncogenes in cancers.

4 Challenges for enhancer prediction

Enhancers are regions of DNA that are responsible for the transcription of one or more genes [46]. The position of enhancers is variable and relative to their target and can occur downstream, upstream, or within the introns of a gene. Enhancers may make 3D contact with promoters to achieve regulation of distal genes [23]. Furthermore, there is no specific motif or code for enhancers, and they may only be active in specific temporal, environmental, and spatial conditions [3]. These characteristics of enhancers make it difficult to identify and annotate the enhancers. The experimental approaches for the identification of enhancers fail to provide a complete list of active enhancers and do not help researchers to understand why certain DNA regions act as enhancers and others do not [5].

5 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is an emerging and necessary field of artificial intelligence, particularly in the field of healthcare, XAI, is designed to enhance human trust in artificial intelligence models by providing the explainability of the model that how a specific model has been generated, and explaining the results of respective models to allow the better understanding of the problem statement. Therefore, it helps the user to improve performance of the model and provide more application domains. Hagraas [19] explains the main important features of XAI including (1) Transparency: It is right to describe how a decision has been made as it affects people's lives, and the explanation should be in a human-understandable format and language. (2) Causality: Does the model provide a complete explanation of the underlying phenomena while providing the correct inferences from the data? (3) Bias: The AI models are trained on the dataset that is coming from the real world, so how can we be sure that these models are also incorporating the biases? (4) Fairness: Can we able to make sure that the decisions that are made by the AI systems are fair? And (5) Safety: Without a depth understanding of the data how can we rely on AI models? A potential XAI system should be able to incorporate all the mentioned features to provide complete transparency to the user.

Hagras [19] also mentions the existing three main approaches to creating an XAI system. The first one is the deep explanation which modifies the deep learning models' techniques to understand explainable structures. A few examples include deepLIFT [39] and layer-wise relevance propagation [7]. The second approach is interpretable models: this is an approach to interpret casual models or learn structures that can be applied to the graphical models, i.e. Hidden Markov Model (HMM), statistical models such as naïve Bayes, logistic regression or random forest. However, the output of these techniques is only understandable by experts and not by laymen. The last approach is the Model Induction: which can be used to interpret the model from any opaque box models. Hagras [19] also mentioned that the best approach to provide the explainability to users is by providing them the IF-THEN rules along with the linguistic labels which can explain model output. Fuzzy logic systems (FLS) are one of the AI technique that provide IF-THEN rules and linguistic labels AI model architecture shown in Fig. 6. FLS has 4 main components 1) Fuzzifier: this converts crisp input into fuzzy sets. 2) Inference: This component generates the ideal rules for respective inputs. 3) Rule base: it contains the membership functions and rules that control or regulate the decision-making process in a fuzzy logic system. The rules are saved here in the form of IF-THEN conditions. And 4) Defuzzification: It transforms the fuzzy set outputs into crisp outputs. An FLS directly converts the real number measurements into linguistic labels. They may take the form of good, fair, bad; high, medium, low, or various combinations of the descriptive variables. Then these linguistic labels are used to define the if-then rule base for describing the situation in a form that is explainable and understandable. An example of a fuzzy rule may be "IF the tumour size is large and the homogeneity is high between the cells THEN the patient has a malignant tumour", here the linguistic labels are large, high, and malignant. It is very simple for any individual, independent of their expertise, to understand what is being measured in the situation and what will be the output. There are two main types of fuzzy logic systems Type-1 FLS and Type-2 FLS. The main distinction between Type-1 FLS and Type-2 FLS is that Type-1 FLS are unable to directly handle the uncertainties because of the specific nature of the membership functions. Type-1 FLS takes the input measured in real numbers also called crisp inputs in terms of fuzzy logic system and fuzzifiers these values into fuzzy sets in the fuzzifier block. After fuzzifying the inputs, the input fuzzy sets map onto the output fuzzy sets by using the fuzzy rules fired in the inference box. Figure 7 represents the Type-1 fuzzy logic membership functions for the decision-making process of early breast cancer detection. Sizilio et al., [40] took two input features (1) tumour area (range: 185–4255) and (2) homogeneity range [0.01–0.45], and the outputs would be either benign (non-cancerous tumour) range [0–0.5], malignant (cancerous tumour) range [0.6–1] or undefined range [0.5–0.6]. In the Type-2 fuzzy logic system instead of defining the crisp membership functions, the fuzzy set includes another representation layer in the form of a

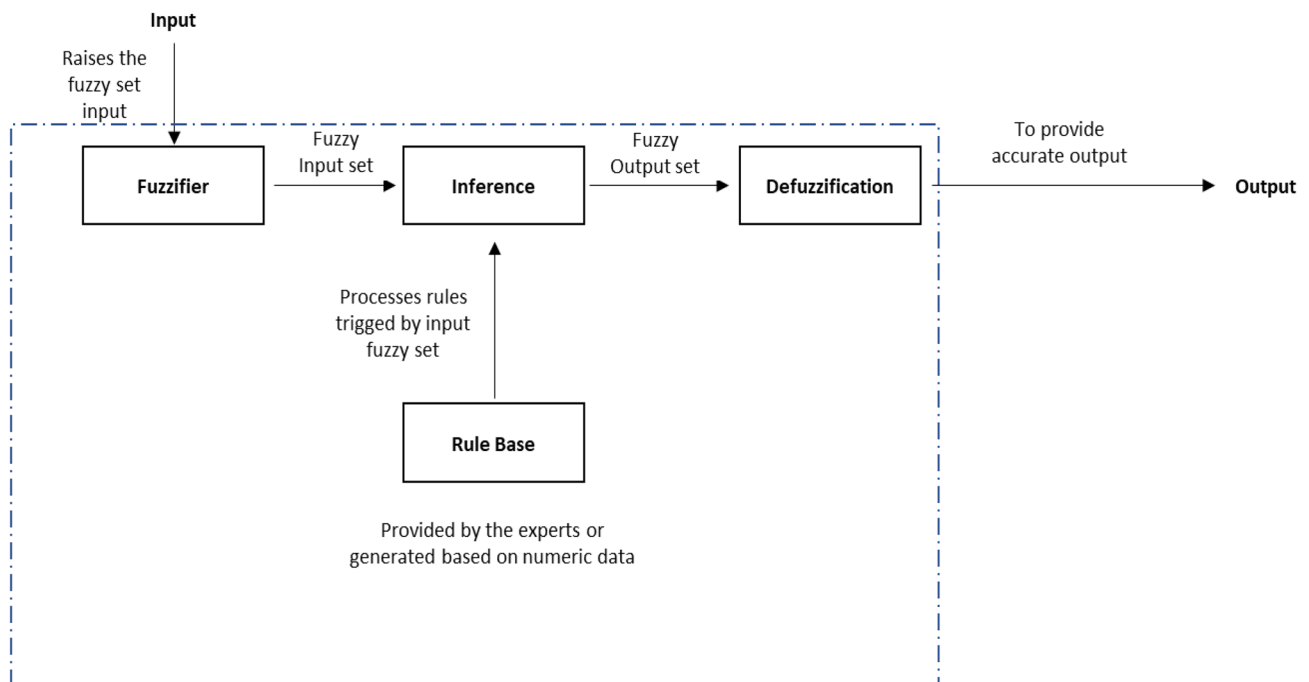


Fig. 6 Systematic Architecture of Fuzzy Logic System (FLS).

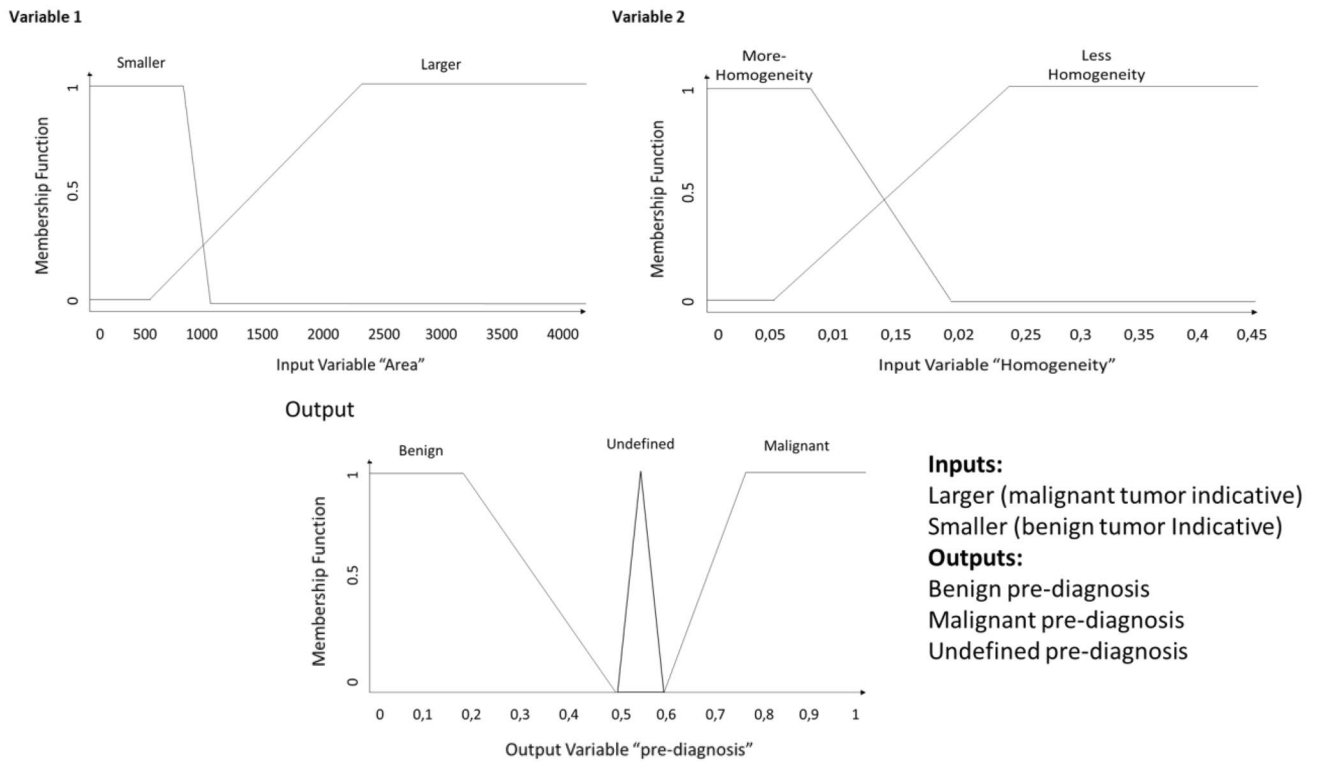


Fig. 7 Representation of the tumour area (smaller and larger), tumour homogeneity (more and less) inputs and Benign, malignant, and undefined output membership functions [36]

footprint of uncertainty (FOU) around the membership functions, which provides the additional degree of freedom to handle the uncertainties [1, 34]. Type-2 inference fuzzy system structure is shown in Fig. 8.

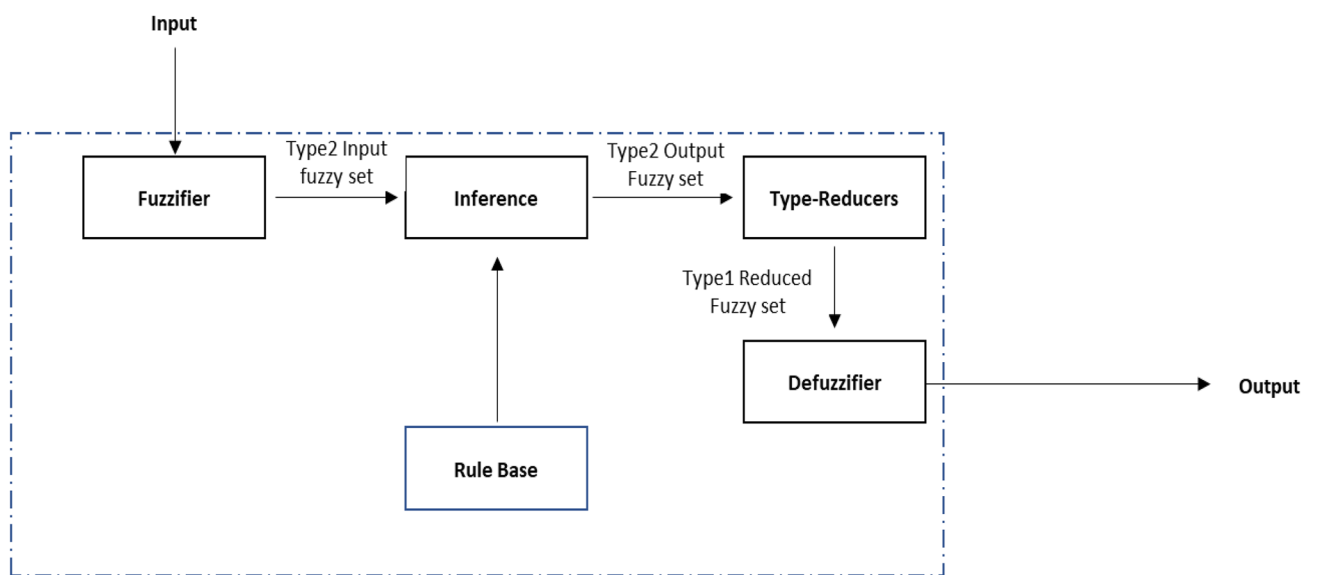


Fig. 8 An overview of the Type-2 Fuzzy Logic System Operation

6 Opaque box AI v/s explainable rule-based models

It is challenging to understand the prediction of the opaque box model (e.g., deep learning) where the cumulative model complexity can be used to achieve high prediction accuracies by these models. Alternatively, interpretable models can provide a better understanding of how predictions have been made. To achieve transparency in the model a concept of explainable AI has been proposed that explains the whole process of the model, i.e. the underlying procedure for explaining the methods, procedures and output of the model that should be understandable by any human [28, 35]. A comparison of the opaque box model and the XAI model is shown in Fig. 9.

The Rule-based explainable AI (XAI) model that generates the natural language IF/THEN rules as a classification algorithm based on type 2 fuzzy logic, generates, integrates, and tests rules for accuracy and validity. This XAI model can help the user to understand which rules are used by the classifier in making the prediction. Rule-based explainable AI (XAI) is a class of artificial intelligence that explain the rule and insights into how AI-based system can make predictions and decisions. XAI can explore the reasoning behind the process of decision-making and provide details on how the system will work in the future and the system's advantages and drawbacks [19]. XAI allows researchers to understand the insights of the predicted results. Opaque-box models like a neural network, random forest and deep learning can always create confusion like "How does the system predict the result", "How does the model work", "Are the results correct", and "How do overcome the errors", "is the result trustworthy". The use of XAI systems can overcome this confusion and provide a clear and transparent prediction with explainable rules[44].

7 Computational methods used for the prediction of cancer

Cancer is a multifaceted and complex disease that continues to be a major healthcare challenge worldwide. Accurate prediction and early detection of cancer are critical but at the same time important for reducing the burden of this disease. In recent years, the advancement in the field of machine learning and artificial intelligence has shown promising results in early detection, diagnosis, and treatment of various cancers. Ström [41] reported that AI in combination with cancer screening methods that include biopsy examination can increase the success rate of breast cancer treatment. Computational radiology uses AI techniques such as computer vision, pattern recognition or lesion detection for the classification of lesions according to Breast Imaging Reporting and Data System (BIRADS) and systematic diagnosis reporting. Mavaddat et al. [25] reported a genetic variant model that calculates the polygenic risk score to estimate the breast cancer risk in a patient. Bakas et al. [8] have proposed a deep convolutional neural network AI-based model that uses magnetic resonance imaging (MRI) data as input and generates rapid and accurate 3D segmentation of glioblastoma. However, the MRI data failed to generate accurate results. Zhou et al. [50] proposed a support vector machine risk model

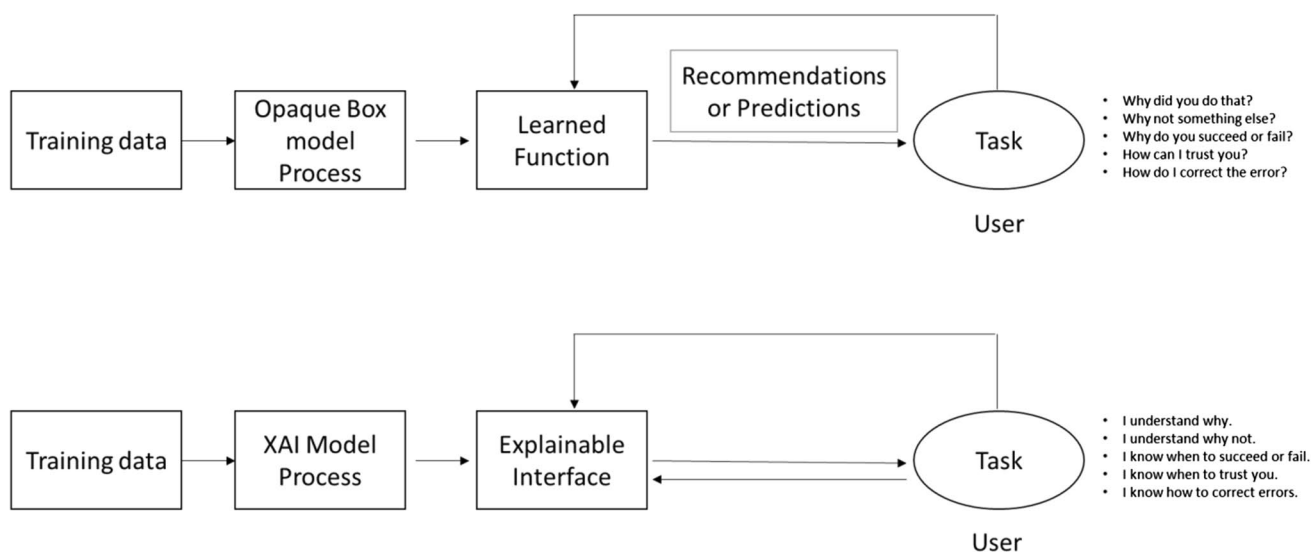


Fig. 9 Comparison between opaque box models and Explainable artificial intelligence

that uses both clinical and genetic data, for predicting ovarian cancer. Mehrotra et al. [27], reported a deep learning-based AI model for the classification of brain tumours. The model is trained on the Magnetic Resonance Imaging (MRI) dataset, and it helps in the classification of both malignant and benign tumour cells. The model achieved an accuracy of 99.04%. Wankhede and Selvarani [47] published an MLL-CNN (multilevel layer model R-CNN) that is based on a relative description model and feature weight factor-based feature selection strategy for the classification of the brain cancer. They trained their model on MRI images to predict the glioblastoma. The model achieves an accuracy of 89%, specificity of 97% and sensitivity of 98%. Toumazis et al. [45] uses the Bayesian model to detect lung cancer based on various risk factors such as smoking exposure, genetics, and age.

However, where the machine learning and Deep learning-based model achieves higher accuracy, these techniques are unable to explain how a particular result has been classified by a specific model. To resolve this issue, Gaur et al. [17] proposed a new model that uses eXplainable AI modelling techniques for the prediction of brain tumours. XAI techniques allow the model to make decisions based on certain rules, that ultimately help researchers or scientists to easily trace results. The study uses the MRI image data for prediction and achieves an accuracy of 94.64%. However, there are still gaps which need to be filled and there is a necessity to develop a molecular level-based feature set for the identification of the real cause of the diseases and their accurate prediction.

8 Enhancers predictions methods

There are numerous experimental techniques used for the identification of enhancers. The first technique is transcription factor binding site mapping onto the genome using ChIP-seq data [38]. The second technique is the use of epigenetic markers (i.e., H3K27ac and H3K4me1) to identify active enhancers. The third approach is the identification of binding sites of the histone acyltransferase EP300, a transcription factor protein that is required for the acetylation of nucleosomes and is recruited by other TFs (Lee & Young [23]). In this approach, the histone modification data will be used to differentiate the active and non-active enhancers [13]. Another approach for genome-wide identification of enhancers is STARR-seq (Self-transcribing active regulatory region sequencing) a massively parallel reporter assay that allows the identification of the enhancers based on the genome-wide activity and provides a quantitative measure of each region in the genome to act as an enhancer and its activity [5].

Computational tools, specifically machine learning tools, are taking the lead in the identification of genome-wide enhancers [36]. These tools use histone modification and high-throughput sequencing assay data as a training data set and based on the extracted features predict the enhancers in genomes. Machine learning methods suffer from biases and tend to predict promoters as enhancers [20]. Promoters are the upstream of Transcription Start Site that define where the RNA polymerase begins the gene transcription [24]. Machine learning methods such as neural networks give high enhancers prediction accuracy, however, they fail to explain the rules and insights through which the algorithm is making the prediction [9]. Additionally, neural networks require large amounts of data as a training set [33].

An accurate prediction of enhancers is necessary to understand the role of non-regulatory genome regions in the context of disease. Wolfe et al. [49] developed a Ruled-based explainable (XAI) model for the identification of the enhancers in *Drosophila melanogaster* cell lines. The model was trained on histone modification ChIP-seq data of histone modifications and STARR-seq data. For evaluating model performance, the XAI model was compared with traditional machine learning models for enhancer prediction and annotation. Using this approach, the machine learning model was trained on the same histone modification data as an explainable model, that accurately predicts enhancer locations and generalises to other cell lines without adjustment. The project was based on the following aims: (1) Train the XAI model on the histone modification ChIP-seq data. (2) Defining, interpreting, and implementing the rules for prediction of the XAI model, and (3) Using this model to predict change in enhancers in other developmental, physiological or disease contexts. A comparison of the opaque box model and the XAI model used for enhancer prediction is shown in Fig. 10.

9 Why explainability is needed

The scientific community is working with a large amount of genomic data, and the focus has shifted to understanding it fully and making it useable for the healthcare sector. The alteration in genomic region of living organisms can cause numerous diseases. Multiple machine learning tools based on neural networks, deep learning and random forests have been developed and have gained high accuracy and efficiency, but these tools lack explainability in their prediction

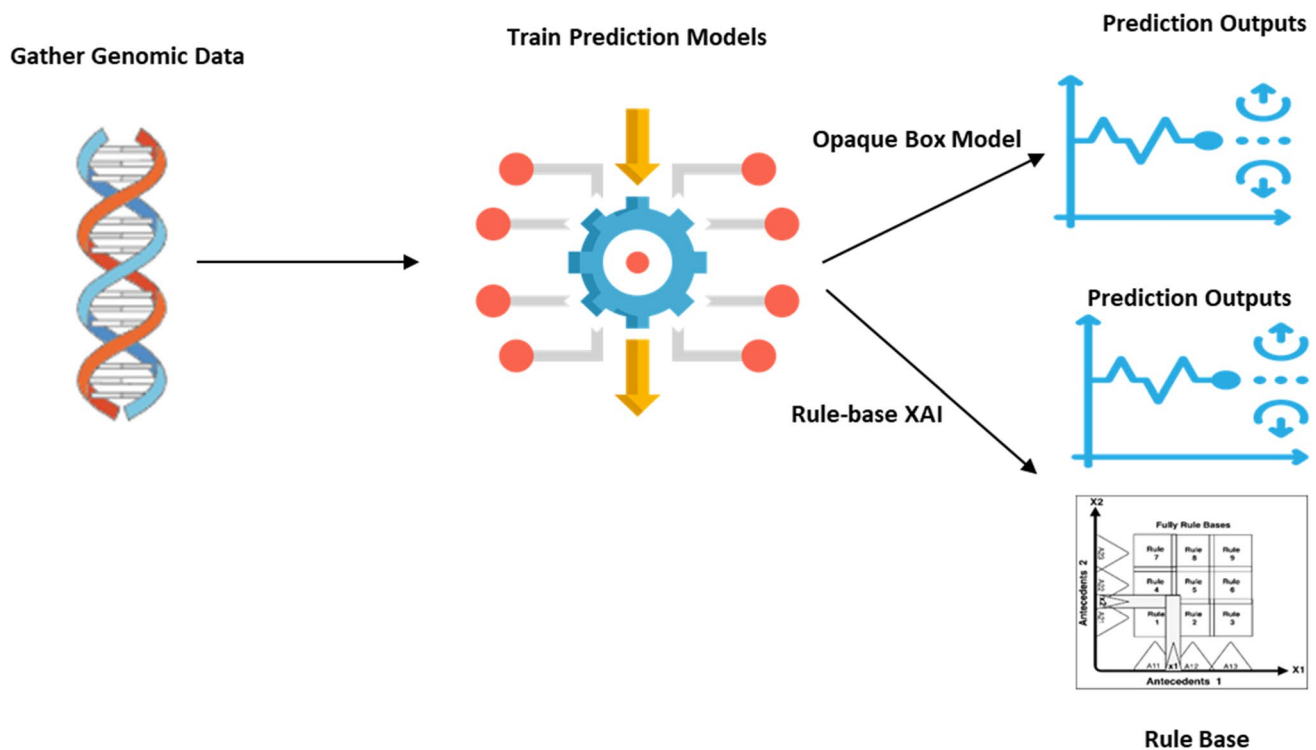


Fig. 10 Comparison of opaque box model and XAI model. Both models are trained on genomic data, The opaque models give non-traceable predictions compared to XAI, which provides predictions along with the IF-THEN rule base that is understandable to layman

results. However, for the genomic scientific community, there is a need to develop explainable generalized models that will help researchers understand the prediction and replicate them clinically to speed up traditional experimental methods aiming to develop new drugs, personalised therapies or propose new treatments and cures for diseases. Therefore, there is a need to offer models that guarantee explainability and transparency in their prediction that will be understandable to a layman which can pave the way to developing predictions quickly to help improve disease outcomes, such as with cancer, through personalized medicines.

Acknowledgements This work was supported by University of Essex (PhD scholarships to K.M). N.R.Z. was supported by Queen Mary University of London. We would like to thank Ines Hofer for comments on the manuscript.

Author contributions K.M., H.H., and N.R.Z. conceived, designed and wrote the paper. The authors read and approved the final manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Acampora G, Alghazawi D, Hagraas H, Vitiello A. An interval type-2 fuzzy logic based framework for reputation management in peer to peer e-commerce. *Inf Sc.* 2016;333:88–107.

2. Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum Genomics*. 2022;16(1):1–20.
3. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhashi E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Müller F, Forrest ARR, Carninci P, Rehli M, Sandelin A. An atlas of active enhancers across human cell types and tissues. *Nature*. 2014;507:455–61. <https://doi.org/10.1038/nature12787>.
4. Arabzadeh A, Mortezaadeh T, Aryafar T, Gharepapagh E, Majdaeen M, Farhood B. Therapeutic potentials of resveratrol in combination with radiotherapy and chemotherapy during glioblastoma treatment: a mechanistic review. *Cancer Cell Int*. 2021;21(1):1–15.
5. Arnold CD, Gerlach D, Stelzer C, Boryn ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*. 2013;339:1074–7. <https://doi.org/10.1126/science.1232542>.
6. Atkinson TJ, Halfon MS. Regulation of gene expression in the genomic context. *Comput Struct Biotechnol J*. 2014;9: e201401001. <https://doi.org/10.5936/csbj.201401001>.
7. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*. 2015;10(7): e0130140.
8. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinohara RT, Berger C, Ha SM, Rozycki M, Prastawa M. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. 2019. <https://doi.org/10.17863/CAM.38755>
9. Calabrese E, Villanueva-Meyer JE, Cha S. A fully automated artificial intelligence method for non-invasive, imaging-based identification of genetic alterations in glioblastomas. *Sci Rep*. 2020;10:11852. <https://doi.org/10.1038/s41598-020-68857-8>.
10. Chathoth KT, Zabet NR. Chromatin architecture reorganization during neuronal cell differentiation in *Drosophila* genome. *Genome Res*. 2019;29(4):613–25.
11. Chen K, Chen Z, Wu D, Zhang L, Lin X, Su J, Rodriguez B, Xi Y, Xia Z, Chen X, Shi X, Wang Q, Li W. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumour-suppressor genes. *Nat Genet*. 2015;47:1149–57. <https://doi.org/10.1038/ng.3385>.
12. Cobb M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol*. 2017;15(9): e2003243.
13. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*. 2010;107:21931–6. <https://doi.org/10.1073/pnas.1016071107>.
14. Daniels H, Jones KH, Heys S, Ford DV. Exploring the use of genomic and routinely collected data: narrative literature review and interview study. *J Med Internet Res*. 2021;23(9): e15739.
15. Del Giacco L, Cattaneo C. Introduction to genomics. In: *Molecular profiling: methods and protocols*. Springer; 2012. p. 79–88.
16. Elliott K, Larsson E. Non-coding driver mutations in human cancer. *Nat Rev Cancer*. 2021;21(8):500–9.
17. Gaur L, Bhandari M, Razdan T, Mallik S, Zhao Z. Explanation-driven deep learning model for prediction of brain tumour status using MRI image data. *Front Genet*. 2022;13:448.
18. Grigorenko EL, Dozier M. Introduction to the special section on genomics. *Child Dev*. 2013;84(1):6–16.
19. Hagras H. Toward human-understandable, explainable AI. *Computer*. 2018;51(9):28–36.
20. Herman-Izycka J, Wlasnowolski M, Wilczynski B. Taking promoters out of enhancers in sequence-based predictions of tissue-specific mammalian enhancers. *BMC Med Genomics*. 2017;10:34. <https://doi.org/10.1186/s12920-017-0264-3>.
21. Karmodiya K, Krebs AR, Oulad-Abdelghani M, Kimura H, Tora L. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics*. 2012;13:424. <https://doi.org/10.1186/1471-2164-13-424>.
22. Kron KJ, Bailey SD, Lupien M. Enhancer alterations in cancer: a source for a cell identity crisis. *Genome Med*. 2014;6(9):1–12.
23. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell*. 2013;152:1237–51. <https://doi.org/10.1016/j.cell.2013.02.014>.
24. Le NQK, Yapp EKY, Nagasundaram N, Yeh HY. (2019). Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext N-grams. *Front Bioeng Biotechnol*. 305.
25. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Maclinnis RJ. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics*, 104(1), 21–34.
26. McGuire AL, Gabriel S, Tishkoff SA, Wonkam A, Chakravarti A, Furlong EE, et al. The road ahead in genetics and genomics. *Nat Rev Genet*. 2020;21(10):581–96.
27. Mehrotra R, Ansari MA, Agrawal R, Anand RS. A transfer learning approach for AI-based classification of brain tumours. *Mach Learn Appl*. 2020;2: 100003. <https://doi.org/10.1016/j.mlwa.2020.100003>.
28. Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: a comprehensive review. *Artif Intell Rev*. 2022;55:1–66.
29. Nam S, Chang HR, Jung HR, Gim Y, Kim NY, Grailhe R, et al. A pathway-based approach for identifying biomarkers of tumour progression to trastuzumab-resistant breast cancer. *Cancer Lett*. 2015;356(2):880–90.
30. Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science*. 2022;376(6588):44–53.
31. Pop RT, Pisante A, Nagy D, Martin PCN, Mikheeva LA, Hayat A, Ficiz G, Zabet, NR. Identification of mammalian transcription factors that bind to inaccessible chromatin. *Nucl Acid Res* 2023;51(16):8480–95. <https://doi.org/10.1093/nar/gkad614>
32. Rothschild MF, Plastow GS. Applications of genomics to improve livestock in the developing world. *Livest Sci*. 2014;166:76–83.
33. Sánchez-Sánchez C, Izzo D. Real-time optimal control via deep neural networks: study on landing problems. *J Guid Control Dyn*. 2018;41(5):1122–1135.
34. Sarabakha A, Imanberdiyev N, Kayacan E, Khanesar M, Hagras, H. Novel Levenberg–Marquardt based learning algorithm for unmanned aerial vehicles. *J Inf Sci*. 2017;417:361–80.

35. Saranya A, Subhashini R. A systematic review of Explainable Artificial Intelligence models and applications: recent developments and future trends. *Decis Anal J.* 2023;7:100230.
36. Sethi A, Gu M, Gumusgoz E, Chan L, Yan K-K, Rozowsky J, Barozzi I, Afzal V, Akiyama JA, Plajzer-Frick I, Yan C, Novak CS, Kato M, Garvin TH, Pham Q, Harrington A, Mannion BJ, Lee EA, Fukuda-Yuzawa Y, Visel A, Dickel DE, Yip KY, Sutton R, Pennacchio LA, Gerstein M. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *Nat Methods.* 2020;17:807–14. <https://doi.org/10.1038/s41592-020-0907-8>.
37. Shang E, Nguyen TTT, Shu C, Westhoff M-A, Karpel-Massler G, Siegelin MD. Epigenetic targeting of Mcl-1 is synthetically lethal with Bcl-xL/Bcl-2 inhibition in model systems of glioblastoma. *Cancers.* 2020;12:2137. <https://doi.org/10.3390/cancers12082137>.
38. Shlyueva D, Stampfel G, Stark A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics,* 15(4), 272–286.
39. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International conference on machine learning. PMLR; 2017. p. 3145–53.
40. Sizilio GR, Leite CR, Guerreiro AM, Neto ADD. Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis. *Biomed Eng Online.* 2012;11(1):1–21.
41. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol.* 2020;21(2):222–32.
42. Teer JK. An improved understanding of cancer genomics through massively parallel sequencing. *Transl Cancer Res.* 2014;3(3):243.
43. Thandapani P. Super-enhancers in cancer. *Pharmacol Ther.* 2019;199:129–38.
44. Tjoa E, Khok HJ, Chouhan T, Cuntai G. (2021). Improving deep neural network classification confidence using heatmap-based eXplainable AI. arXiv preprint: <https://arXiv.org/abs/2201.00009>.
45. Toumazis I, Bastani M, Han SS, Plevritis SK. (2020). Risk-based lung cancer screening: a systematic review. *Lung Cancer,* 147, 154–186.
46. Tung YA, Yang WT, Hsieh TT, Chang YC, Wu JT, Oyang YJ, Chen CY. accuEnhancer: Accurate enhancer prediction by integration of multiple cell type data with deep learning. 2020. <https://doi.org/10.1101/2020.11.10.375717>
47. Wankhede DS, Selvarani R. Dynamic based architecture-based deep learning approach for glioblastoma brain tumour survival prediction. *Neurosci Inf Artif Intell Brain Inf.* 2022;2: 100062. <https://doi.org/10.1016/j.neuri.2022.100062>.
48. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediators establish super-enhancers at key cell identity genes. *Cell.* 2013;153(2):307–19.
49. Wolfe JC, Mikheeva LA, Hagrais H, Zabet NR. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in *Drosophila*. *Genome Biol.* 2021;22:308. <https://doi.org/10.1186/s13059-021-02532-7>.
50. Zhou J, Li L, Wang L, Li X, Xing H, Cheng L. (2018). Establishment of a SVM classifier to predict recurrence of ovarian cancer. *Molecular Medicine Reports,* 18(4), 3589–3598.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.