Unravelling the application of machine learning in cancer biomarker discovery

Carter William [a, *], Choki Wangmo [a], Anjali Ranjan [b]

[a] *James Cook University, Brisbane, Australia*

[b] *Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India*

ABSTRACT

Machine learning is playing an increasingly important role in the healthcare industry by transforming the way cancer is diagnosed and treated. By analyzing patient data, genomic data, and imaging data, machine learning algorithms can identify molecular signatures that distinguish cancer patients from healthy patients. Biomarkers that can accurately detect and diagnose cancer can be identified through analysis of these data sources. Additionally, personalized cancer therapies can be developed by identifying the most effective treatments based on individual patient characteristics and cancer type. Some of the machine learning techniques used for cancer biomarker discovery include deep learning and support vector machines, which can respectively identify complex patterns in data and classify data to identify relevant biomarkers. The benefits of using machine learning for cancer biomarker discovery are significant, including more precise and personalized treatments, improved patient outcomes, and the potential to transform cancer diagnosis and treatment. However, there are also challenges associated with using machine learning for cancer biomarker discovery, such as data collection and privacy issues, as well as the need for more powerful computational resources. This article explores the potential of machine learning in cancer biomarker discovery and argues that ongoing research in this field has the potential to revolutionize cancer diagnosis and treatment. Future research directions should focus on further developing machine learning algorithms and effective data collection and privacy protocols.

KEYWORDS

Cancer; Biomarker; Machine learning; Artificial intelligence; Bioinformatics

*Corresponding author: Carter William
E-mail address: anjaliranjan914@gmail.com

1. Introduction

Cancer biomarkers are molecules found in the body that can be utilized to identify a specific type of cancer or provide information about the tumor. [1][2]. The identification of cancer biomarkers is a rapidly expanding field, and machine learning (ML) is playing an increasingly vital role in this endeavor. ML, a type of artificial intelligence (AI) that employs data and algorithms to learn from experience, is being utilized extensively. [3]. ML algorithms have the ability to detect patterns in vast datasets and are being utilized to identify potential cancer biomarkers. By analyzing large datasets, including genomic data, ML can detect even subtle differences between cancerous and healthy cells. This makes it possible to identify novel biomarkers for cancer. [4]. ML can also be utilized to analyze existing cancer biomarkers and assist in the identification of biomarkers that are more precise and reliable. Such an approach can aid doctors in accurately diagnosing and treating cancer patients. [5]. As an example, ML algorithms can be employed to analyze existing cancer biomarkers and establish which ones are most effective in detecting a specific type of cancer or predicting treatment outcomes. [4]. Aside from biomarker discovery, ML can also enhance the precision of cancer diagnosis and treatment. ML algorithms can scrutinize extensive datasets, detect patterns that predict the presence of cancer, and improve the accuracy of diagnosis and treatment. [6]. Overall, ML is playing an increasingly vital role in the discovery of cancer biomarkers. Through analyzing massive datasets, ML can discern even subtle variations between cancerous and healthy cells, which can lead to the identification of potential cancer biomarkers. Furthermore, ML can refine the accuracy of cancer diagnosis and treatment. These advances in ML have the potential to significantly impact the way cancer is diagnosed and treated in the future. [7][8].

2. Role of ML in theranostics of cancer

Recently, ML has been employed in theranostic studies for cancer, which encompass the diagnosis, prognosis, treatment, and management of cancer. [9]. Theranostic studies are a fusion of diagnostics and therapeutics, which aid in disease diagnosis, identifying the optimal treatment, and enhancing the patient's prognosis. ML has facilitated the evolution of personalized medicine, whereby treatments are customized to suit each individual patient's requirements. [10][11]. ML plays a crucial role in such studies by identifying the most effective treatments and enhancing diagnosis accuracy. ML is currently employed in identifying biomarkers for cancer diagnosis and prognosis. [12]. **Fig. 1** depicts a representative flowchart for identifying potential cancer biomarkers. Biomarkers are biological indicators that can reveal the presence or risk of disease. ML algorithms can detect patterns in data to identify potential biomarkers. Moreover, ML can analyze imaging data, such as MRI scans, to highlight areas of concern and facilitate treatment identification. [13]. In addition, ML has emerged as a crucial tool in theranostic studies for cancer, which aim to diagnose, prognose, treat, and manage cancer. ML plays a key role in identifying the most effective treatment and improving the accuracy of diagnosis. One area where ML is being used is in identifying potential biomarkers for cancer diagnosis and prognosis. A biomarker is a biological indicator that can be measured to indicate the presence or risk of a disease. By analyzing large datasets, ML algorithms can identify patterns in the data that can be used to identify potential biomarkers. ML can also be used to analyze medical imaging data, such as MRI scans, to identify areas of concern and potential treatments. Moreover, ML is also being used to develop new treatments for cancer. ML algorithms can be used to analyze the data from clinical trials and identify the most effective treatments. ML can also be used to identify potential new drugs that may be effective in treating cancer. Furthermore, ML is being used to improve radiotherapy treatments by analyzing patient data and identifying potential areas that may benefit from radiotherapy. ML algorithms can also identify the best radiation dose and the most effective radiation delivery methods. Finally, ML is also

improving the accuracy of cancer diagnosis and prognosis by analyzing patient data and identifying potential factors that may influence the outcome of the disease. ML can analyze a patient's genetic data to improve the accuracy of diagnosis and prognosis. Overall, ML has made a profound impact on theranostic studies for cancer, enabling the development of personalized medicine, the identification of biomarkers, and the improvement of treatment options. [14].
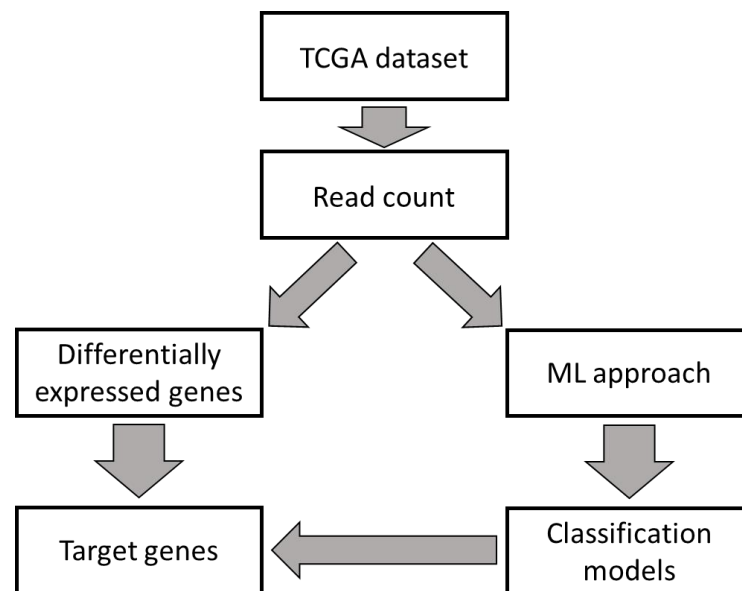

Figure 1. A representative flowchart showing a representative flowchart for the identification of novel biomarkers for cancer, with TCGA referring to The Cancer Genome Atlas.

3. Different types of ML models for classification and prediction in cancer studies

ML algorithms have various applications in cancer research, including the classification of different types of cancer, detection of cancerous regions in medical images, and prediction of the prognosis for cancer patients. [15]. Here, we will discuss various types of classification and prediction models for cancer studies using ML algorithms (**Fig. 2**). The most commonly used ML classification algorithm for cancer studies is the supervised learning algorithm. [16]. This algorithm uses labeled data to learn how to classify different types of cancer. The data can include patient demographic information, lab results, medical imaging results, and gene expression data. The algorithm can then use this data to classify cancer into various types such as breast cancer, lung cancer, or prostate cancer. Such classification is crucial for diagnosing and treating cancer, as it can provide doctors with more detailed information about the cancer and how to treat it. Another type of classification and prediction model is the unsupervised learning algorithm. This algorithm uses unlabeled data to identify patterns and clusters in the data without the need for labels. This is useful for identifying and predicting cancer risks in a population, as it can detect patterns in the data that may indicate an increased risk for cancer. [17][18]. The third type of classification and prediction model for cancer studies using ML algorithms is the deep learning algorithm. Deep learning algorithms are used to develop more complex models for cancer studies. These models use a large amount of data to create a detailed representation of the cancer, which can be used for diagnosis, prognosis, and treatment purposes. Deep learning algorithms are particularly effective in processing complex medical imaging data, such as CT scans, MRI scans, and X-rays. These algorithms can identify subtle patterns in the images that may not be visible to the human eye, and use this information to make more accurate predictions about the cancer.

[19]. ML algorithms are becoming increasingly important in cancer research, enabling the development of powerful tools for classification and prediction. In **Fig. 2**, we illustrate the different types of ML algorithms used for cancer study. One of the most commonly used algorithms is supervised learning, which employs labeled data to classify cancer into different types, such as breast, lung, or prostate cancer. This information is crucial for diagnosis and treatment planning. Another type of algorithm is unsupervised learning, which can identify patterns and clusters in unlabeled data. This is useful for predicting cancer risk in populations. A third type is deep learning, which is used to create complex models that can identify patterns in images and provide detailed representations of cancer for diagnosis, prognosis, and treatment. Finally, predictive models can be used to forecast the outcome of a patient's treatment based on individual characteristics and medical history. These models can help clinicians choose the most appropriate treatment for each patient. Overall, ML algorithms are versatile and powerful tools that have revolutionized cancer research and have the potential to improve patient outcomes. [20][21].
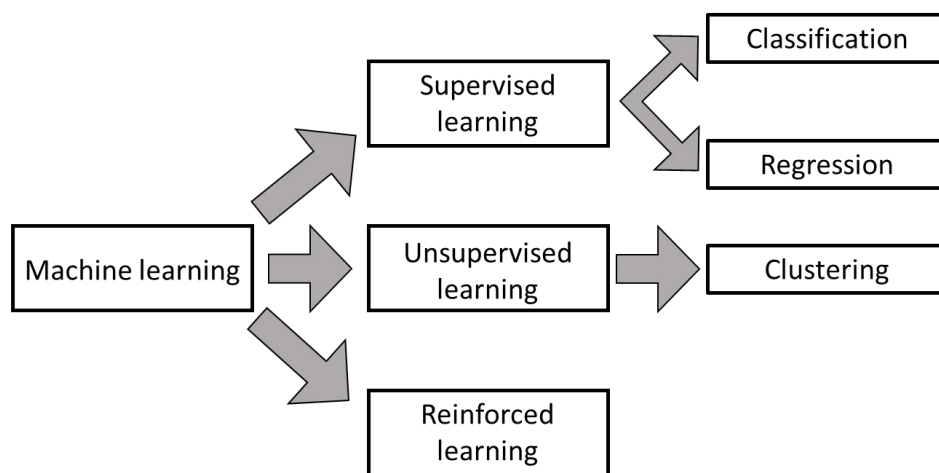

Figure 2. Main types of ML models. Supervised learning involves approaches like classification and regression, while unsupervised learning involves clustering. Reinforcement learning is a type of ML that involves an agent interacting with its environment to improve the model's performance.

4. Drawbacks of ML algorithm for cancer diagnosis

ML algorithms have made a significant breakthrough in the field of medical diagnosis, enabling medical professionals to diagnose diseases more quickly and accurately than ever before. However, there are some limitations to using ML algorithms for cancer diagnosis. One of the major limitations is the input data used to train the algorithm. If there is insufficient training data available to properly train the algorithm, it will not be able to make accurate predictions. [22]. Additionally, if the training data is biased, then the algorithm may make incorrect predictions. This can be especially dangerous in cancer diagnosis, as inaccurate diagnoses can lead to inappropriate treatments or missed opportunities for early interventions. Thus, it is important to ensure that the input data used for training the algorithm is diverse and representative of the population to avoid biases. Furthermore, ML algorithms should not be considered a replacement for human expertise and clinical judgment. Instead, they should be used as a supplementary tool to assist medical professionals in making more informed decisions. It is crucial to strike a balance between the benefits of ML algorithms and the expertise of medical professionals to ensure the best possible patient outcomes. [23]. Second, Despite the advantages of ML algorithms in medical diagnosis, there are some limitations to their use in cancer diagnosis. Firstly, these algorithms heavily rely on the quality and quantity of input data, and the lack of proper training data can hinder their ability to make

accurate predictions. Moreover, if the training data is biased or unrepresentative of the actual population, the algorithm may generate incorrect results, which can be dangerous in the case of cancer diagnosis. Secondly, ML algorithms can be slow and inefficient in processing data, potentially delaying the diagnosis process, which can be critical for cancer treatment. Finally, the amount of data that these algorithms can process is limited, and an insufficient number of samples may impact their ability to accurately diagnose the type of cancer, while an excessive number may overwhelm the algorithm, leading to inaccurate diagnoses. [24]. Overall, ML algorithms have been a significant breakthrough in the field of medical diagnosis, allowing for faster and more accurate diagnosis of diseases such as cancer. However, there are some limitations to their use. Firstly, the accuracy of ML algorithms relies heavily on the quality and quantity of input data, and if there is not enough training data available or if the data is biased, it can lead to incorrect predictions. This can be particularly risky when dealing with cancer diagnosis as a wrong diagnosis can lead to improper treatment or missed opportunities for early intervention. Secondly, ML algorithms can be slow and inefficient in processing the data, which can lead to delays in diagnosis, and in cancer diagnosis, timely diagnosis is crucial for successful treatment. Thirdly, ML algorithms have limitations in processing vast amounts of data accurately. If there is too little data, the algorithm may not be accurate in diagnosis, while too much data can overwhelm the algorithm and make it less efficient. Finally, ML algorithms are not always foolproof and may make mistakes, particularly when presented with new or unknown data or when applied to a different population than the one on which it was trained. Therefore, medical professionals must be aware of these limitations when deciding whether to use ML algorithms for cancer diagnosis. [25].

5. Does ML tool need doctors for making medical decision?

The use of ML tools in the medical field has seen significant growth in recent years and continues to evolve. However, it is important to note that while ML tools have the potential to revolutionize the medical field, they cannot replace the expertise of medical professionals. Medical professionals play a crucial role in the successful application of ML tools in medicine. While the tools can provide valuable information and data, it is ultimately the responsibility of the medical professionals to interpret the data and make informed decisions. Medical professionals are required to have a deep understanding of the implications of the data, as well as the implications of the decisions they make. Therefore, it is essential to ensure that medical professionals receive proper training and education on how to effectively use ML tools in medicine, as well as how to interpret the results to make the best decisions for their patients. [26][27]. For example, while a ML tool may be capable of detecting anomalies in a patient's health data, it cannot explain why the anomaly exists or what it means for the patient's health. Likewise, ML tools cannot provide suggestions for how to improve the patient's health. Medical professionals play a crucial role in ensuring the accuracy and impartiality of the data collected by ML tools. They can verify that the data is appropriately collected and that it is not biased in any way. Additionally, medical professionals can ensure that the data is correctly interpreted and that the decisions based on the data are suitable. Furthermore, medical professionals are essential for implementing ML tools in the medical field. They can provide guidance to guarantee that the tools are used in an ethical and responsible manner. Medical professionals can also provide feedback on the performance of the tools and suggest areas for improvement. In conclusion, ML tools have the potential to revolutionize the medical field, but they require the expertise and guidance of medical professionals to be successfully applied. Medical professionals are vital for interpreting the data, making informed decisions, and ensuring the accuracy and ethical use of the tools. [28][29].

6. Serious and harmful effects of ML in medical decision

ML is a powerful tool in medical decision-making, but it can also be dangerous if not used properly. Careful consideration and oversight are needed to ensure that ML algorithms are used responsibly. There are several examples of dangerous effects of ML in medical decision-making, including: (i) Over-reliance on algorithms: While ML algorithms can be accurate, they are not perfect and can make mistakes. Relying solely on algorithm-generated results can lead to wrong or even dangerous decisions. For example, an algorithm might incorrectly diagnose a patient with a condition they don't have or recommend a treatment that is not appropriate for their condition. (ii) Bias and discrimination: ML algorithms can be biased if the data used to train them is biased. This can lead to unfair or discriminatory decisions. For instance, an algorithm trained on data from predominantly white patients may be more likely to recommend treatments that are more beneficial for them than for other racial groups. (iii) Unintended consequences: ML algorithms can have unexpected and unintended consequences. For instance, an algorithm trained on data from a specific population may not work well when applied to a different population. This can lead to wrong decisions and unexpected outcomes. (iv) Security concerns: Data used for training ML algorithms can be vulnerable to hacking, leading to wrong decisions and data breaches if the data is stolen or tampered with. (v) High cost: ML algorithms can require a lot of computing power and can be expensive to develop and maintain, making it difficult for medical institutions and other organizations to use them. These are just some of the potential risks of using ML in medical decision-making. To mitigate these risks, it's important to use data that is unbiased and of high quality, implement appropriate security measures, and have systems in place to monitor and evaluate the performance of the algorithms. In addition, medical professionals should provide oversight and interpret the results generated by the algorithms to ensure the decisions made are appropriate and safe for patients. [30][29].

7. Conclusion and future perspectives

ML is a subfield of AI that uses algorithms to learn from and make decisions based on data. In recent years, ML has had a transformative impact on biomedical sciences, playing a critical role in advancing clinical care by providing automated decision support and aiding in the diagnosis of diseases with greater speed and accuracy. In the field of biomedical sciences, ML is used in numerous areas, such as drug discovery, personalized medicine, and medical imaging. For instance, ML algorithms are employed to analyze large datasets to identify potential drug targets, uncover new therapeutic strategies, and optimize drug delivery systems. In addition, ML is also used to analyze patient data to identify disease subtypes or predict patient responses to a particular drug therapy. By automating the interpretation of medical images like CT scans and X-rays, ML algorithms help speed up the diagnosis process. Clinical science has also benefited from ML, which can identify high-risk patients, predict patient responses to a particular treatment, and automate the analysis of medical images and electronic health records (EHR) to enhance accuracy and reduce diagnosis time. ML algorithms can also detect subtle patterns in EHRs that could indicate developing disorders like sepsis or kidney failure. In the future, ML has the potential to revolutionize medical research, clinical practice, and patient care. For example, ML algorithms could be used to automate the analysis of genetic data and provide insights into the underlying causes of various diseases. Furthermore, ML can help develop personalized treatments tailored to individual patients, which reduces healthcare costs and improves patient outcomes. Lastly, ML can be used to automate clinical trial design and enhance the accuracy of clinical trial results. In summary, ML has transformed biomedical and clinical sciences, and the technology has the potential to revolutionize medical research and patient care. By offering automated decision support and identifying subtle patterns in data, ML expedites the diagnosis process and improves patient outcomes.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Carter William; Writing–original draft：Carter William, Choki Wangmo; Writing–review & editing: Anjali Ranjan. All authors read and approved the final manuscript.

References

1. Henry, N.L.; Hayes, D.F. Cancer biomarkers. *Mol. Oncol.* **2012**, *6*, 140–146, doi:10.1016/j.molonc.2012.01.010.

2. Sarhadi, V.K.; Armengol, G. Molecular Biomarkers in Cancer. *Biomolecules* **2022**, *12*, 1021, doi:10.3390/biom12081021.

3. Shao, D.; Dai, Y.; Li, N.; Cao, X.; Zhao, W.; Cheng, L.; Rong, Z.; Huang, L.; Wang, Y.; Zhao, J. Artificial intelligence in clinical research of cancers. *Brief. Bioinform.* **2022**, *23*, doi:10.1093/bib/bbab523.

4. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M. V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17, doi:10.1016/j.csbj.2014.11.005.

5. Farina, E.; Nabhen, J.J.; Dacoregio, M.I.; Batalini, F.; Moraes, F.Y. An overview of artificial intelligence in oncology. *Futur. Sci. OA* **2022**, *8*, doi:10.2144/fsoa-2021-0074.

6. Koh, D.-M.; Papanikolaou, N.; Bick, U.; Illing, R.; Kahn, C.E.; Kalpathi-Cramer, J.; Matos, C.; Martí-Bonmatí, L.; Miles, A.; Mun, S.K.; et al. Artificial intelligence and machine learning in cancer imaging. *Commun. Med.* **2022**, *2*, 133, doi:10.1038/s43856-022-00199-0.

7. Tran, K.A.; Kondrashova, O.; Bradley, A.; Williams, E.D.; Pearson, J. V.; Waddell, N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med.* **2021**, *13*, 152, doi:10.1186/s13073-021-00968-x.

8. Wang, S.; Wang, S.; Wang, Z. A survey on multi-omics-based cancer diagnosis using machine learning with the potential application in gastrointestinal cancer. *Front. Med.* **2023**, *9*, doi:10.3389/fmed.2022.1109365.

9. Chen, Z.; Lin, L.; Wu, C.; Li, C.; Xu, R.; Sun, Y. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Commun.* **2021**, *41*, 1100–1115, doi:10.1002/cac2.12215.

10. Jeelani, S.; Jagat Reddy, R.; Maheswaran, T.; Asokan, G.; Dany, A.; Anand, B. Theranostics: A treasured tailor for tomorrow. *J. Pharm. Bioallied Sci.* **2014**, *6*, 6, doi:10.4103/0975-7406.137249.

11. Johnson, K.B.; Wei, W.; Weeraratne, D.; Frisse, M.E.; Misulis, K.; Rhee, K.; Zhao, J.; Snowdon, J.L. Precision Medicine, AI, and the Future of Personalized Health Care. *Clin. Transl. Sci.* **2021**, *14*, 86–93, doi:10.1111/cts.12884.

12. Hunter, B.; Hindocha, S.; Lee, R.W. The Role of Artificial Intelligence in Early Cancer Diagnosis. *Cancers (Basel).* **2022**, *14*, 1524, doi:10.3390/cancers14061524.

13. Hajjo, R.; Sabbah, D.A.; Bardaweel, S.K.; Tropsha, A. Identification of Tumor-Specific MRI Biomarkers Using Machine Learning (ML). *Diagnostics* **2021**, *11*, 742, doi:10.3390/diagnostics11050742.

14. Krzyszczyk, P.; Acevedo, A.; Davidoff, E.J.; Timmins, L.M.; Marrero-Berrios, I.; Patel, M.; White, C.; Lowe, C.; Sherba, J.J.; Hartmanshenn, C.; et al. The growing role of precision and personalized medicine for cancer treatment.

TECHNOLOGY **2018**, *06*, 79–100, doi:10.1142/S2339547818300020.

15. Cruz, J.A.; Wishart, D.S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2007**, *2*, 59–77.

16. Uddin, S.; Khan, A.; Hossain, M.E.; Moni, M.A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 281, doi:10.1186/s12911-019-1004-8.

17. Lynch, C.M.; van Berkel, V.H.; Frieboes, H.B. Application of unsupervised analysis techniques to lung cancer patient data. *PLoS One* **2017**, *12*, e0184370, doi:10.1371/journal.pone.0184370.

18. Javaid, M.; Haleem, A.; Pratap Singh, R.; Suman, R.; Rab, S. Significance of machine learning in healthcare: Features, pillars and applications. *Int. J. Intell. Networks* **2022**, *3*, 58–73, doi:10.1016/j.ijin.2022.05.002.

19. Ebrahim, M.; Sedky, A.A.H.; Mesbah, S. Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer. *Data* **2023**, *8*, 35, doi:10.3390/data8020035.

20. Buda, M.; Saha, A.; Walsh, R.; Ghate, S.; Li, N.; Swiecicki, A.; Lo, J.Y.; Mazurowski, M.A. A Data Set and Deep Learning Algorithm for the Detection of Masses and Architectural Distortions in Digital Breast Tomosynthesis Images. *JAMA Netw. Open* **2021**, *4*, e2119100, doi:10.1001/jamanetworkopen.2021.19100.

21. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310, doi:10.1186/s12911-020-01332-6.

22. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160, doi:10.1007/s42979-021-00592-x.

23. Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann. Intern. Med.* **2018**, *169*, 866, doi:10.7326/M18-1990.

24. Gianfrancesco, M.A.; Tamang, S.; Yazdany, J.; Schmajuk, G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern. Med.* **2018**, *178*, 1544, doi:10.1001/jamainternmed.2018.3763.

25. Pai, R.K.; Van Booven, D.J.; Parmar, M.; Lokeshwar, S.D.; Shah, K.; Ramasamy, R.; Arora, H. A review of current advancements and limitations of artificial intelligence in genitourinary cancers. *Am. J. Clin. Exp. Urol.* **2020**, *8*, 152–162.

26. Davenport, T.; Kalakota, R. The potential for artificial intelligence in healthcare. *Futur. Healthc. J.* **2019**, *6*, 94–98, doi:10.7861/futurehosp.6-2-94.

27. Ahuja, A.S. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ* **2019**, *7*, e7702, doi:10.7717/peerj.7702.

28. Giordano, C.; Brennan, M.; Mohamed, B.; Rashidi, P.; Modave, F.; Tighe, P. Accessing Artificial Intelligence for Clinical Decision-Making. *Front. Digit. Heal.* **2021**, *3*, doi:10.3389/fdgth.2021.645232.

29. Sanchez-Martinez, S.; Camara, O.; Piella, G.; Cikes, M.; González-Ballester, M.Á.; Miron, M.; Vellido, A.; Gómez, E.; Fraser, A.G.; Bijnens, B. Machine Learning for Clinical Decision-Making: Challenges and Opportunities in Cardiovascular Imaging. *Front. Cardiovasc. Med.* **2022**, *8*, doi:10.3389/fcvm.2021.765693.

30. McDonald, L.; Ramagopalan, S. V.; Cox, A.P.; Oguz, M. Unintended consequences of machine learning in medicine? *F1000Research* **2017**, *6*, 1707, doi:10.12688/f1000research.12693.1.