ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 10, Issue 5, May 2024

MACHINE LEARNING-BASED CANCER DETECTION: A SYSTEMATIC REVIEW

[1] Moolya Chaitra Satish
Student
Alva's Institute of Engineering and Technology
lavanyakulal777@gmail.com

[2] Mutturaj Unki
Student
Alva's Institute of Engineering and Technology
mutturajunki01@gmail.com

[3] Mydam Niharika
Student
Alva's Institute of Engineering and Technology
mydamniharikashetty@gmail.com

[4] Najmul Huda
Student
Alva's Institute of Engineering and Technology
najmulhuda2812@gmail.com

[5] Brahma Prakash H P
Associate Professor
Alva's Institute of Engineering and Technology
drbrahmap@aiet.org.in

ABSTRACT

The advancement of genome sequencing technology has made it possible for researchers to think outside the box and imaginatively[1]. Researchers are putting a lot of effort into fighting certain hereditary diseases, such as cancer. Artificial intelligence has made medical research more powerful. The accessibility of publicly available healthcare statistics has incentivized academics to create applications that facilitate prompt disease detection and prognosis. Recent strides in artificial intelligence, particularly the integration of state-of-the-art Deep Learning (DL) structures with conventional Machine Learning (ML) techniques, have ushered in a transformative era for medical oncology and cancer research[24]. In this extensive review study, we highlight important studies, approaches, and results while offering a summary of the most current ML applications in cancer research. Using a focus on publications from the previous five years, the chosen papers were found using the PubMed and dblp databases. Our findings led to the classification of the many uses of ML in cancer research into three main clinical scenarios. The goal of this review is to further knowledge about the changing field of machine learning in cancer and how it may affect patient outcomes.

I. INTRODUCTION

Cancer is not a single disease, but rather a collection of related conditions involving uncontrolled cell division and proliferation. It ranks second in the developing world and takes the lives of over 8 million people annually in the industrialized world. Early cancer diagnosis and prognosis are becoming critical components of cancer research because they can aid in the following clinical management of patients. Accurately distinguishing benign from malignant tumors is essential for improving clinical decision-making. The classification of cancer into high-risk and low-risk categories has typically been done using statistical methods, despite the complex linkages found in high-dimensional medical data[19].

Recent advances in cancer prognosis and prediction have made use of machine learning to overcome the shortcomings of conventional statistical methods. Machine learning is a branch of artificial intelligence that makes use of various statistical, probabilistic, and optimization techniques to allow computers to "learn" from past examples and recognize challenging patterns from large, noisy, or complex data sets. This feature is particularly well suited for applications in medicine, especially those that require complex proteomic and genomic data. Because of this, machine learning is frequently used in the diagnosis and detection of cancer. This latter

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 10, Issue 5, May 2024

strategy is especially intriguing because it fits with the expanding trend of personalized and predictive treatment. The field of bioinformatics is proving to be indispensable in the fight against a number of life-threatening illnesses, including diabetes, Alzheimer's, and cancer. Mutations and changes in an individual's genetic microenvironment are what lead to cancer[4].

Treatment is difficult because of the complexity of the cancer microenvironment. Even if two patients have the same form of cancer, their responses to the same kind of therapy will differ. Clinical trials and the conventional drug research process are tedious and time-consuming procedures. As a result, scientists are putting a lot of effort into creating the best medicines they can for such challenging circumstances. The fight against diabetes, Alzheimer's disease, cancer, and other life-threatening disorders is showing the value of bioinformatics.AI has recently made its way into the scientific research of several diseases, including cancer, clinical practice, and translational medicine. Current artificial intelligence systems, which just use machine learning methods, are being used in many therapeutic settings. These areas include: (i) image-based computer-aided diagnosis and detection in various medical specialties, such as pathology, radiology, ophthalmology, and dermatology; (ii) interpreting genomic data to identify genetic variants using high-throughput sequencing technologies; (iii) prognostic and monitoring patient information; (iv) finding new biomarkers by combining omics and phenotype data; (v) determining health status based on biological signals gathered from wearable devices; and finally (vi) developing and utilizing autonomous robots in medical interventions[6].

II. CATEGORIZATION OF CANCER PATIENTS

Based on well-established machine learning methods for addressing binary or multi-class learning problems, the classification problem of sickness prediction in medical oncology and cancer research has been thoroughly addressed. By grouping patients into predetermined categories, machine learning (ML)-based predictive models capable of evaluating risk stratification with broadly applicable performance might be developed. Regarding this, a number of research studies that used DL techniques and conventional algorithms to forecast the identification of critical variables for cancer classification were published last year. The bulk of studies employed DL architectures to assess imaging and genomic data in order to forecast and stratify risks. Evidently, DL models were trained using genetic and imaging data to identify and categorise illness subgroups[8].

The results of studies demonstrate that cancer disorders are more like large disease families with a multitude of sub-types, and that categorizing tumors according to their anatomical features is not as appropriate as classifying them according to the pathological alteration of signaling pathways at the cellular level. This distinction is crucial because, although a given treatment may be completely relevant and successful for one patient, it may have no beneficial effect on tumour control and simply have adverse consequences in other people with the "same" cancer[11]. A robust statistical foundation and evidence-based medicine depend greatly on the quantity and calibre of available data. The more pertinent aspects there are, the more data is needed.

In order to help generalise biological processes and systems, machine learning primarily focuses on finding hidden patterns in data. Improving the classification model's prediction accuracy and identifying the smallest possible group of putative gene biomarkers are the two main goals of cancer classification.

III. RESEARCH APPROACHES

There are several approaches to the recognition, prediction, and classification of various cancer kinds. The following processes, along with databases, feature extraction techniques, preprocessing techniques, and classification approaches, are the main ways that machine learning is used to identify cancer. The processes involved in the ML and DL algorithms used for the detection and segmentation of the various types of cancer include feature extraction, preprocessing, classification, and database management.

The automated diagnosis and detection of cancer is unquestionably one of the most important and productive fields of biomedical machine learning applications. Previous study papers suggested machine learning (ML)-based pipelines based on traditional or cutting-edge methods to carry out diagnostic tasks in common cancer kinds like breast, lung, colon, and pancreatic cancers, among others.

Most studies created automated diagnostic models primarily employing DL architectures and imaging

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 10, Issue 5, May 2024

data from positron-emission tomography (PET), computed tomography (CT), magnetic resonance imaging (MRI), and X-ray radiography.

1. DATA COLLECTION AND PREPROCESSING

Utilizing reputable sources such as TCGA and SEER, the focus is placed on acquiring datasets with a diverse range and high-quality annotations. For private datasets, adherence to permission protocols is mandatory[31].

The subsequent step in the data cleaning process involves meticulously removing noise and unrelated information from healthcare records and medical imaging. Augmentation techniques such as image cropping and rotation are used to overcome the problem of insufficient data and improve the model's ability to generalize.

Concurrently, standardization protocols are implemented to normalize image pixel values and standardize clinical data. This ensures uniformity across sources. To further improve the overall quality of the data, missing values are corrected using appropriate imputation techniques.

2. FEATURE EXTRACTION

Feature extraction begins with the use of deep learning architectures, specifically CNNs, to extract pertinent information from medical images. To increase the model's ability to generalize to various cancer kinds, transfer learning from pre-trained models, such as Inception or ResNet, is being researched. Clinical data integration also involves extracting relevant information from patient records, such as genetic information, medical history, and demographics. Methods that consider the temporal aspects of the data are explored for merging clinical and image-based variables. Using Principal Component Analysis and other dimensionality reduction approaches, the number of dimensions in the model is reduced, improving computation speed[13].

3. MODEL DEVELOPMENT

To capture complex patterns in cancer data, the development of the AI model necessitates careful consideration of deep learning architectures, such as CNNs, RNNs, and ensemble models[30]. Ensuring optimal performance involves fine-tuning model parameters through methods like grid search or Bayesian optimization in hyper parameter optimization[26].

The incorporation of transfer learning from pre-trained models is key to leveraging knowledge from large-scale datasets, thereby enhancing the model's generalization ability. Additionally, pre-processing methods, including batch normalization and dropout, are investigated. These methods serve to improve model convergence and mitigate over fitting[35].

4. TRAINING AND VALIDATION

To test model performance robustly, the model is evaluated on many dataset subsets using K-fold cross-validation. Over fitting during training is avoided by a well stated termination condition, and convergence is ensured by constant monitoring. AUC-ROC, F1-score, sensitivity, specificity, accuracy, and other comprehensive performance indicators offer a detailed assessment of the model's efficacy[16].

5. POST-PROCESSING AND ERROR ANALYSIS

Tailored to the clinical context, post-processing involves balancing sensitivity and specificity by optimizing decision thresholds. Utilizing the confusion matrix, a comprehensive error analysis identifies prevalent misclassifications and patterns, guiding further refinement of the model.

6. IMPLEMENTATION AND DEPLOYMENT

Collaborating with experts to integrate healthcare systems ensures a seamless incorporation into current workflows and compliance with legal requirements like HIPAA. Ongoing monitoring methods and timely upgrades, informed by new information and research, ensure the model's relevance in real-world circumstances[35].

7. ETHICAL DISCUSSIONS

Protecting patient privacy through robust anonymization techniques and strictly following informed consent rules are two ethical considerations. It is imperative to tackle biases in training data and predictions, with a specific emphasis on reducing demographic inequalities to ensure impartial and equitable results[7].

IV. CONCLUSION

We emphasized the fundamentals of machine learning in this review and how they relate to cancer prognosis and prediction. Most of the current research has been devoted to developing prediction

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 10, Issue 5, May 2024

models using supervised machine learning techniques and classification algorithms in order to forecast illness outcomes accurately. Examining their results makes it evident that multidimensional heterogeneous data integration, in conjunction with different approaches to feature selection and classification, might yield valuable inference tools for the field of cancer research[29].

REFERENCES

[1] Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719–31.

[3] Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. J R Soc Med 2019;112(1):22–8.

[4] M. Chen and M. Decary, "Artificial intelligence in healthcare: An essential guide for health leaders," in Healthcare management forum, 2020, pp. 10-18.

[5] Huang S, Yang J, Fong S, Zhao Qi. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer Lett 2020;471:61–71.

[6] Azuaje F. Artificial intelligence for precision oncology: beyond patient stratification. NPJ Precis Oncol 2019;3:1–5.

[7] Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and Engineering Ethics*, 24(2), 505–528.

[8]Li, S., Wu, X., & Tan, M. (2008). Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Computing*, 12(11), 1039–1048.

[9]Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3784-3797.

[10] Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JPA, Vaux DL. Reproducible research practices and transparency across the biomedical literature. PLoS Biol 2016;14(1):e1002333.

[11]Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinf 23(19):2507–2517

[12] Inza I, Larrañaga P, Blanco R, Cerrolaza AJ (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. ArtifIntell Med 31(2):91–103

[13] Shen Qi, Shi W-M, Kong W (2008) Hybrid particle swarm optimization and tabu search approach for selecting genes for tumorclassifcation using gene expression data. Comput Biol Chem 32(1):53–60

[14] Li S, Xixian Wu, Tan M (2008) Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Comput 12(11):1039–1048

[15]Branke J, Deb K, Dierolf H, Osswald M (2004) Finding knees in multi-objective optimization. International conference on parallel problem solving from nature. Springer, Berlin, Heidelberg, pp 722–731

[16] Marler RT, Arora JS (2004) Survey of multi-objective optimization methods for engineering. Struct MultidiscipOptim 26(6):369–395

[17] BoussaïD I, Lepagnot J, Siarry P (2013) A survey on optimization metaheuristics. Inf Sci 237:82–117

[18] Chakraborty A, Kar AK (2017) Swarm intelligence: a review of algorithms. In: Nature-inspired computing and optimization. Springer, pp 475–494

[19] Weinberg RA (1991) Tumor suppressor genes. Science 254(5035):1138–1146

[20] Knoechel B, Roderick JE, Williamson KE, Zhu J, Lohr JG, Cotton MJ, Gillespie SM (2014) An epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemi. Nat Genet 46(4):364–370

[21] I. Kumar, B. H.s., J. Virmani, S. Thakur, 2017. A classification framework for prediction of breast density using an ensemble of neural network classifiers, Biocybernetics and Biomedical Engineering. 37, 217–228. doi:10.1016/j.bbe.2017.01.001.

[22] H.H. Aghdam, E.J. Heravi. 2017. Convolutional Neural Networks, Guide to Convolutional Neural Networks, 85–130. doi:10.1007/978-3-319-57550-6_3.

[23] Z. Jiao, X. Gao, Y. Wang, J. Li. 2017. A parasitic metric learning net for breast mass classification based on mammography. Pattern Recognition. doi:10.1016/j.patcog.2017.07.008.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 10, Issue 5, May 2024

[24] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv. Neural Inform. Process. Syst. 2012;25:1097–105.

[25] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput1997;9:1735–80.

[26] Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. Neural Comput2000;12:2451–71.

[27] Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge University Press; 2014.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, pp. 436- 444, 2015.

[29] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.

[30] Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. in Proceedings of the international conference on machine learning, 2013.

[31]Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483(7391):603–607

[32] Yamada M, Lian W, Goyal A, Chen J, Wimalawarne K, Khan SA, Chang Y (2017) Convex factorization machine for toxicogenomics prediction. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1215–1224

[33] Jamali M, Ester M (2010) A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on Recommender systems, pp 135–142

[34] Wang L, Li X, Zhang L, Gao Q (2017) Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer 17(1):513–524

[35] Evans WE, McLeod HL (2003) Pharmacogenomics drug disposition, drug targets, and side efects. N Engl J Med 348(6):538–549

[36] Wei D-Q, Wang J-F, Chen C, Li Y, Chou K-C (2008) Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. Protein Pept Lett 15(1):27–32 4892 A Systematic Review of Applications of Machine Learning in Cancer Prediction and Diagnosis 1 3

[37] Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y (2012) Relating drug–protein interaction network with drug side efects. Bioinformatics 28(18):i522–i528

[38] Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast fexible docking method using an incremental construction algorithm. J Mol Biol 261(3):470–489

[39] Xie Li, Evangelidis T, Xie L, Bourne PE (2011) Drug discovery using chemical systems biology weak inhibition of multiple kinases may contribute to the anti-cancer efect of nelfnavir. PLoSComput Biol 7(4):e1002037

[40] Jacob L, Vert J-P (2008) Protein-ligand interaction prediction an improved chemogenomics approach. Bioinformatics 24(19):2149–2156