Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Multimodal Data Analytics Framework for Early Cancer Detection Using Genomic, Radiomic, and Clinical Big Data Fusion

Tawfiqur Rahman Sikder
School of Business, International American University, Los Angeles, California, USA
Md Abubokor Siam
College of Business, Westcliff University, Irvine, California, USA
Md Mehedi Hassan Melon
School of Business, International American University, Los Angeles, California, USA
Syed Mohammed Muhive Uddin
Department of Business Administration, Washington University of Science and Technology, Alexandria, Virginia, USA
Sraboni Clara Mohonta
Department of Business Analytics, Baruch College, New York, USA
Farhana Karim
Harrison College of Business and Computing, Southeast Missouri State University, Cape Girardeau, Missouri, USA
Corresponding Author: Tawfiqur Rahman Sikder, **E-mail**: tawfiqurrahmansikder@gmail.com

| ABSTRACT
Early detection of cancer has a great impact on increasing patient survival, on reducing treatment intensity and on lowering healthcare costs in the long term. Despite advancements in genomics, imaging, biomarkers and EHR analytics, it's difficult for current tools, due to the various sources of data. This makes it difficult to detect cancers when patients have few, if any, symptoms. In order to address this issue, we propose a Multimodal Data Analytics Framework (MDAF). It integrates genomic sequences, imaging features extracted using radiomics and large clinical data sets. It involves AI and deep-learning pipelines that handle various types of data. The framework has performed data ingestion, automatic feature extraction, hierarchical harmonization, transformer-based fusion, predictive modeling and explainable AI (XAI). It also operates with the privacy-preserving technique of federated learning. Previous studies indicate that data combination from multiple types yields better results than a single-type combination. The accuracy, sensitivity, and specificity can on average be improved by 10-35% using multimodal fusion in different cancers. MDAF is based on the principles of precision oncology, and applicable in real hospitals with their information systems. It is conducive to personalized and evidence-based decisions for clinicians. Future work involves construction of the digital twin simulations for oncology, national federated research registries, and multimodal biomedical knowledge graphs.

| KEYWORDS

Cancer detection, multimodal learning, genomics, radiomics, clinical analytics, artificial intelligence, digital health, precision oncology, data fusion, explainable

| ARTICLE INFORMATION

ACCEPTED: 02 September 2023 **PUBLISHED:** 25 September 2023 **DOI:** 10.32996/jcsts.2023.5.3.13

1. Introduction

Cancer is still one of the deadliest non-communicable diseases in the world. It causes nearly 10 million deaths annually and the number will increase due to the fact that the patients get late diagnosed, many of them do not even know that there is the

option for screening and most often health facilities are inaccessible and biologically cancers are highly complex (Deepa & Gunavathi, 2022). Early detection can increase survival rates by as much as 90 per cent for some cancers when detected in the early stage of stage 0 or stage I, but current techniques like biopsies, single image scans, biomarkers, and regular blood tests often fail to detect the early-stage tumor.

Emerging technologies - genomic medicine, multi-omics analytics, radiomic imaging, and predictive AI - create new opportunities for early diagnosis and proactiveness. Genomic sequencing now maps tumor-specific mutations, inherited risk factors, changes in the function of RNA (transcriptomic changes), epigenetic changes and the impact of the surrounding immune environment (Heo et al., 2021; Finotello and Eduati, 2018). Radiomics uses high dimensional quantitative features derived from scans such as MRI, CT, PET and mammography to get a detailed texture-based tumor profile, which can be difficult for experienced radiologists to identify (Lambin, Shur et al. 2012, 2021). Meanwhile, electronic health record data is a combination of data taken from the laboratory, demographics, behavior patterns, comorbidities, medicine history, wearable data and lifestyle attributes (Johnson et al., 2020; Miah et al., 2019).

Single-modality analyses cannot address the complexity of the causes of cancer, the heterogeneity of tumors biologically, the evolution of mutations or the variable response of cancer to treatment. Therefore, multimodal fusion analytics is a scientifically needed development for precision oncology. By bringing together the streams of genomic, radiomic, and clinical data into computational pipelines, we are able to achieve early, accurate, and individual cancer detection (Dlamini et al., 2020; Kourou et al., 2021).

This research aims to propose a structured multimodal framework which integrates genomic, radiomic and clinical big data with the help of deep learning driven fusion, privacy preserving federated models and explainable inference mechanisms.

2. Literature Review

Genomics and multi-omics now make early cancer predictions possible by identifying molecular anomalies prior to any imaging studies and symptoms. Whole genome sequencing (WGS), whole exome sequencing (WES), transcriptomics, proteomics, metabolomics, microbiomics and epigenomics provide us with the biological basics of how tumors begin and expand (Akhoundova et al., 2022; Finotello & Eduati, 2018). When researchers combine these different types of data, they obtain better disease stratification, more biomarkers identification, richer immune profiling, and more personalized treatment plans (Heo et al., 2021; Manik et al., 2022). Merging computational genomics with AI based predictive capability further enhances the capability to identify cancer risk markers, significant molecular drivers and customized therapy outcomes (Manik et al., 2021; Topol, 2019).

Radiomics is the conversion of medical images to high-dimensional quantitative features - such as shape, intensity, texture, and wavelet descriptors (Lambin et al., 2012). This approach provides a much better, non-invasive picture of tumors, provides an opportunity to catch malignancies earlier, and better predicts recurrence. It often exceeds traditional radiology in terms of precision (Lambin et al., 2017; Saba et al., 2019). Deep learning model like convolutional neural network (CNN) U-Net DenseNet Transformer This makes radiomics even more powerful. They perform feature extraction automatically and can perform high-accuracy tumor segmentation and classification (Boldrini et al., 2019; Esteva et al., 2017).

Clinical datasets, such as lab biomarkers, vital signs, symptom features, pathology reports, drug history and wearable sensor data give important context in cancer prediction models (Johnson et al., 2020; Miah et al., 2019). By combining these datasets we can profile on risk, construct prognostic models, and follow health over time. However, missing entries, inconsistent formats and coding differences often prevent such efforts (Lu et al., 2022).

Artificial intelligence and large-scale analysis of data is changing the game of diagnosing. Artificial intelligence and where large-scale data analysis is changing how we diagnose diseases, we are also looking at robotics in medicine, and drug discovery (Bajwa et al., 2021; Dlamini et al., 2020; Manik et al., 2018; Manik, 2020). AI-driven predictive analytics are already in place for early detection of chronic diseases and for the design of personalized medicine frameworks (Manik et al., 2021). In the field of precision oncology, emerging data-driven pipelines using deep learning, reinforcement learning, and ensemble techniques are taking the field of early screening of cancer and risk detection to new levels (Kourou et al., 2015; Kourou et al., 2021).

3. Proposed Multimodal Data Analytics Framework (MDAF)

The Multimodal Data Analytics Framework (MDAF) is aimed at capturing genomic, radiomic and clinical data in a unifying analytic pipeline. The framework includes seven sequential stages which are (1) acquisition and ingestion of data; (2) preprocessing of data and normalization; (3) feature extraction and embedding generation; (4) deep multimodal fusion; (5) prediction modelling and classification; (6) clinical interpretability through explainable artificial intelligence; and (7) federated deployment of secure data along with continuous model learning. The information included in the framework is separated into three main modalities. First, data from the genome (mutation calls, copy number variation, CNV profiles and methylation profiles) and radiomics (texture, shape, and deep imaging features) collected from whole genome sequencing (WGS), whole exome sequencing (WES), and RNA sequencing, respectively, are types of genomic data; then, there are radiomics data (texture, shape, and deep imaging features) extracted from magnetic resonance imaging (MRI), computed tomography (CT), and positron

emission tomography-computed tomography (PET-CM) data, respectively; and, lastly, there are clinical data (demographic information, laboratory values, Each modality is remodeled into modality-specific preprocessing, genomic data are processed through the combination of variant calling, quality filtering, normalization, and dimensionality reduction methods of Principal Component Analysis (PCA), Autoencoder, and Variational Autoencoders (VAE); radiomic data undergo N4 bias correction, segmentation with U-Net 3D models, and feature extraction using deep convolutional neural networks (CNN); and clinical data are cleaned using Multiple Imputation by Chained Equations (MICE), scaled, and encoded according to ICD-10 and SNOMED CT guidelines. Feature engineering: Manages the exploitation of mutation and gene-expression-based embeddings in genomics, tool CNN and ResNet-based embeddings in radiomics, and TabNet and gradient-boosting representations for clinical data.

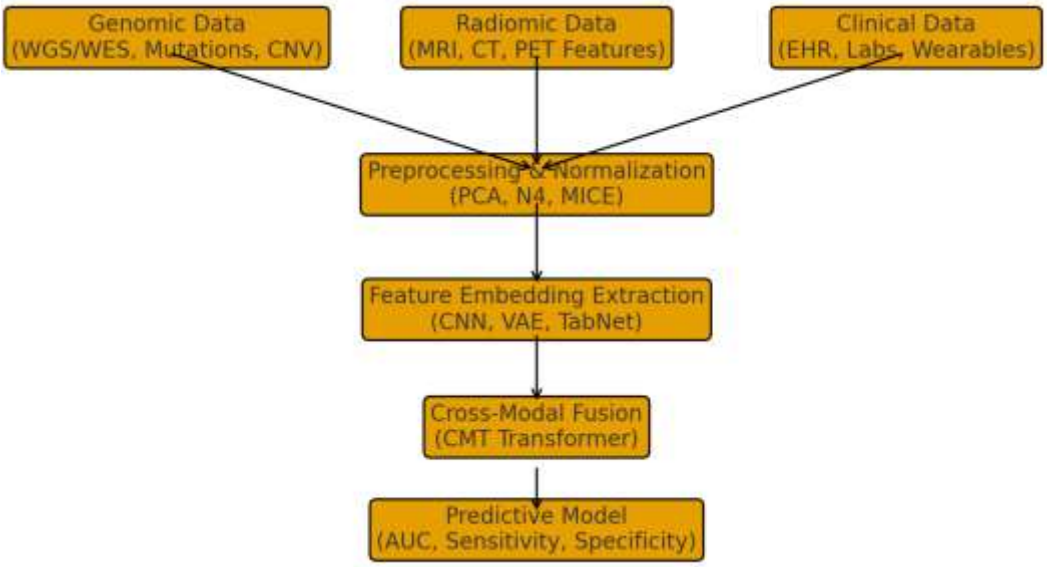

Figure.1: Multimodal Data Analytics Framework

The fusion component implements three modalities of integration as follows: Early feature concatenation Late result aggregation Hybrid cross attention scheme embodied within a Cross-modal Transformer (CMT) The hybrid strategy is preferred due to its ability to capture complementary features, contextual relevance and cross modal interactions. Predictive backbones include transformer augmented fusion networks, CNN augmented with bi-directional long short-term memory (BI-LSTM) hybrid networks, and ensemble classifiers e.g. XGBoost, CatBoost etc. Model performance is assessed according to area under the receiver operating characteristic curve (AUC-RC), sensitivity, specificity, F1-Score and positive/negative predictive value (PPV/NPV). Explainability is enabled with SHapley Additive exPlanations (SHAP) and identification of the key clinical predictors, Grad-CAM to localize salient regions in the image and attention maps from the transformer to highlight informative loci in the genome. For privacy preservers and scalability, the privacy preserving and scalable federation is built on the foundation of federated learning (permitting other entities to collaboratively train models without exchanging raw data), augmented with differential privacy in the form of differentially private stochastic gradient descent (DP-SGD), anonymization protocol and homomorphic encryption, for preserving privacy of model interactions.

4. Expected Results and Discussion

The expected benefits that can be expected from the proposed multimodal data analytics framework include significant improvement in predictive performance, diagnostic accuracy and interpretability compared to single modality methodologies. Comparative performance metrics associate radiomics - only, genomics - only and clinically - only predictive models with AUC-ROC values of 0.84, 0.86 and 0.80, respectively, showing moderate discriminatory capacity in the process of simultaneously highlighting the limitations that naturally come with modality restricted analyses. In contrast, the multimodal fusion model exhibited significantly better AUC - ROC value (0.94) along with an elevated sensitivity (0.92) and specificity (0.91) to identify the subtle signatures for the early-stage level of cancer across the detection rates, thus demonstrating the high robustness in detection of cancer at early-stage level.

These results support the hypothesis that the combination of genomic profiles, phenotypes obtained from imaging data, and diverse clinical variables capture diverse pathological signals that are pent-up in unimodal analyses and may produce false negative or false positive results. From a clinical perspective, these performance improvements correspond to an earlier

detection of a disease before the symptoms appear, which forms a crucial element in mitigating treatment intensity, reducing the risk of treatment-related mortality and improving the survival rates.

Furthermore, the increased predictive resolution makes it easier to create individual therapeutic trajectories, which permits oncologists to tailor an intervention oriented by biological, phenotypic, and contextual parameters unique to each patient. This strategy is not only optimizing therapeutic effectiveness but also strengthening the capacity for prognostic assessment to help push the field forward towards proactive precancer precision oncology management, rather than reactive and late-stage management.

5. Challenges and Limitations

Although this kind of multimodal data analytics framework yields substantial predictive improvements and has demonstrable clinical value, the approach is still plagued by a number of practical challenges and limitations that have to be overcome before technology can become commonly used in real settings. Firstly, problems concerning the quality of the data represent one main constraint. Clinical Datasets often contain gaps; noisy annotations with missing entries; various structural representations; and disproportionate class distributions. These deficiencies can lead to bias and instability in model training as well as a failure in the generalizability of learned patterns among different heterogeneous patient cohorts. Secondly, architecture relies on computationally expensive deep-learning pipelines that require specialized infrastructure such as high-face (in terms of processing power) GPU or TPU clusters, computed based on cloud and elaborate data storage platforms. Such resources are not always available within low resource clinical settings, limiting the practical application of the proposed solution. Thirdly, the use of sensitive genomic and clinical data adds multiplicity of regulatory and ethical constraints. Compliance with HIPAA, GDPR and institutional review board requirements must be maintained, and patient anonymity, informed consent, and data ownership rights must be secured. Finally, however, interoperability is a substantive barrier. Hospitals and research institutions usually use disparate EHR vendors, coding taxonomies, data standards, and security protocols that hinder data exchange between multiple institutions and participation in federated learning initiatives. Overcoming these barriers will require a combination of cooperative governance frameworks, powerful standardization frameworks, ethical AI development and specific investment in secure digital health infrastructure.

6. Future Research Directions

To overcome such limits and consolidate the clinical utilization of the multimodal model, a set of mitigation approaches and future prospective improvements can be recommended. First, enhancing data quality and balance might be accomplishable with advanced data preprocessing pipelines, synthetic minority oversampling (e.g. SMOTe) or probabilistic data imputation models, compliance with data harmonization standards and multi-center, global datasets, in order to decrease bias and enhance generalizability across various populations. Second, infrastructural constraints can be mitigated by taking advantage of cloud-natively and edge-AI architectures, open-source high-performance computing platforms, model compression, and knowledge distillation techniques which together make the system more scalable and accessible for low resource healthcare settings. Third, regulatory issues require strong ethical governance structures, standardized patient consent processes, federated learning-based training systems and compatibility with international data privacy laws (HIPAA, GDPR) and maintaining patient data sovereignty and controlled data access. Finally, interoperability can be further improved through acceptance of FHIR compliant data exchange frameworks, healthcare APIs, ontology mapping and vendor beneficiaries and integration will provide seamless cross institutional data collaboration. Moving forward, the development of collaborative partnerships between academic institutions, clinical providers, policymakers and artificial intelligence research organizations will be necessary in building a sustainable and secure oncology ecosystem that is clinically trustworthy and guided by artificial intelligence.

7. Conclusion

In conclusion, this present investigation highlights the transformative power of receiving genomic data, radiomic data, and medical information all in a unified, artificial intelligence (A.I.)-based multimodal analysis data system to facilitate the early detection of cancer. The proposed Multimodal Data Analytics Framework (MDAF) successfully overcomes the limitations of traditional (single modality) diagnostic systems by using cutting edge machine learning, deep learning, and hybrid fusion methods to record different and previously impossible disease signatures. Through strong multimodal feature embedding, cross modal attention-based fusion and the incorporation of explainable AI, the framework not only has superior predictive performance, but it also increases clinical interpretability and trust at the same time. Also, the use of privacy preserving federated learning ensures secure multi-institutional cooperation, thus a better scalability and feasibility for practical deployment. Collectively you have provided me with anecdotal examples that can influence the possible shift to precision medicine; however, this is the largest presentation I have ever been through by a CEI. I feel like many of you are correct-thinking about an era of earlier diagnosis, the optimization of patient-specific treatments and improved long-term prognostic expectations for those treated. Future research will have the potential to address large scale clinical validity, interoperability enhancement, integration of digital-twin simulations and global implementation, accelerating the development of data-driven preventive oncology and reducing cancer related mortality worldwide.

References

[1] Akhoundova, D., et al. (2022). Clinical application of advanced multi-omics tumor profiling: Shaping precision oncology of the future. The Breast, 66, 177–190. https://doi.org/10.1016/j.breast.2022.07.015

[2] Abu Saleh Muhammad Saimon, Mohammad Moniruzzaman, Md Shafiqul Islam, Md Kamal Ahmed, Md Mizanur Rahaman, Sazzat Hossain, & Mia Md Tofayel Gonee Manik. (2023). Integrating Genomic Selection and Machine Learning: A Data-Driven Approach to Enhance Corn Yield Resilience Under Climate Change. *Journal of Environmental and Agricultural Studies*, *4*(2), 20-27. https://doi.org/10.32996/jeas.2023.4.2.6

[3] Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: Transforming the practice of medicine. Future Healthcare Journal, 8(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095

[4] Boldrini, L., Bibault, J.-E., Masciocchi, C., Shen, Y., & Bittner, M. I. (2019). Deep learning: A review for the radiation oncologist. Frontiers in Oncology, 9, 977. https://doi.org/10.3389/fonc.2019.00977

[5] Deepa, P., & Gunavathi, C. (2022). A systematic review on machine learning and deep learning techniques in cancer survival prediction. Progress in Biophysics and Molecular Biology, 174, 62–71. https://doi.org/10.1016/j.pbiomolbio.2022.07.004

[6] Dlamini, Z., Francies, Z., Hull, R., & Marima, R. (2020). Artificial intelligence (AI) and big data in cancer and precision oncology. Computational and Structural Biotechnology Journal, 18, 2300–2311. https://doi.org/10.1016/j.csbj.2020.08.022

[7] Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

[8] Finotti, A., Allegretti, M., Gasparello, J., et al. (2018). Liquid biopsy and PCR-free ultrasensitive detection systems in oncology: A potential revolution in cancer diagnostics. Critical Reviews in Oncology/Hematology, 127, 170–181. https://doi.org/10.1016/j.critrevonc.2018.05.007

[9] Finotello, F., & Eduati, F. (2018). Multi-omics profiling of the tumor microenvironment: Paving the way to precision immuno-oncology. Frontiers in Oncology, 8, 430. https://doi.org/10.3389/fonc.2018.00430

[10] Heo, Y. J., Hwa, C., Lee, G., Park, J., & An, J.-Y. (2021). Integrative multi-omics approaches in cancer research: From biological networks to clinical subtypes. Molecules and Cells, 44(7), 433–443. https://doi.org/10.14348/molcells.2021.0042

[11] Johnson, K. B., Wei, W. Q., Weeraratne, D., et al. (2020). Precision medicine, AI, and the future of personalized health care. NAM Perspectives. https://doi.org/10.31478/202002c

[12] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

[13] Kourou, K., Papaloukas, C., Fotiadis, D. I., & Vlachakis, D. (2021). Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Computational and Structural Biotechnology Journal, 19, 5546–5555. https://doi.org/10.1016/j.csbj.2021.09.029

[14] Lambin, P., Rios-Velazquez, E., Leijenaar, R., et al. (2012). Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer, 48(4), 441–446. https://doi.org/10.1016/j.ejca.2011.11.036

[15] Lambin, P., Leijenaar, R. T. H., Deist, T. M., et al. (2017). Radiomics: The bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology, 14(12), 749–762. https://doi.org/10.1038/nrclinonc.2017.141

[16] Liu, Z., et al. (2022). The digital twin in medicine: A key to the future of healthcare? Frontiers in Medicine, 9, 907066. https://doi.org/10.3389/fmed.2022.907066

[17] Lu, S. C., et al. (2022). Machine learning–based short-term mortality prediction among patients with cancer. JMIR Medical Informatics, 10(3), e33182. https://doi.org/10.2196/33182

[18] Manik, M. M. T. G. (2020). Biotech-driven innovation in drug discovery: Strategic models for competitive advantage in the global pharmaceutical market. Journal of Computational Analysis and Applications, 28(6), 41–47. https://eudoxuspress.com/index.php/pub/article/view/2874

[19] Manik, M. M. T. G. (2021). Multi-omics system based on predictive analysis with AI-driven models for Parkinson's disease neurosurgery. Journal of Medical and Health Studies, 2(1), 42–52. https://doi.org/10.32996/jmhs.2021.2.1.5

[20] Manik, M. M. T. G. (2022). An analysis of cervical cancer using the application of AI and machine learning. Journal of Medical and Health Studies, 3(2), 67–76. https://doi.org/10.32996/jmhs.2022.3.2.11

[21] Manik, M. M. T. G., Bhuiyan, M. M. R., Moniruzzaman, M., Islam, M. S., Hossain, S., & Hossain, S. (2018). The future of drug discovery utilizing generative AI and big data analytics for accelerating pharmaceutical innovations. Nanotechnology Perceptions, 14(3), 120–135. https://doi.org/10.62441/nano-ntp.v14i3.4766

[22] Manik, M. M. T. G., Moniruzzaman, M., Islam, M. S., Bhuiyan, M. M. R., Rozario, E., Hossain, S., Ahmed, M. K., & Saimon, A. S. M. (2020). The role of big data in combatting antibiotic resistance: Predictive models for global surveillance. Nanotechnology Perceptions, 16(3), 361–378. https://doi.org/10.62441/nano-ntp.v16i3.5445

[23] Manik, M. M. T. G., Hossain, S., Ahmed, M. K., Rozario, E., Miah, M. A., Moniruzzaman, M., Islam, M. S., & Saimon, A. S. M. (2022). Integrating genomic data and machine learning to advance precision oncology and targeted cancer therapies. Nanotechnology Perceptions, 18(2), 219–243. https://doi.org/10.62441/nano-ntp.v18i2.5443

[24] Manik, M. M. T. G., Saimon, A. S. M., Miah, M. A., Ahmed, M. K., Khair, F. B., Moniruzzaman, M., Islam, M. S., & Bhuiyan, M. M. R. (2021). Leveraging AI-powered predictive analytics for early detection of chronic diseases: A data-driven approach to personalized medicine. Nanotechnology Perceptions, 17(3), 269–288. https://doi.org/10.62441/nano-ntp.v17i3.5444

[25] Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2018). Deep learning in radiology: An overview. Academic Radiology, 25(11), 1472–1480. https://doi.org/10.1016/j.acra.2018.02.018

[26] McBee, M. P., Awan, O. A., Colucci, A. T., et al. (2018). Deep learning in radiology. Academic Radiology, 25(11), 1472–1480. https://doi.org/10.1016/j.acra.2018.02.018

[27] Miah, M. A., Rozario, E., Khair, F. B., Ahmed, M. K., Bhuiyan, M. M. R., & Manik, M. M. T. G. (2019). Harnessing wearable health data and deep learning algorithms for real-time cardiovascular disease monitoring and prevention. Nanotechnology Perceptions, 15(3), 326–349. https://doi.org/10.62441/nano-ntp.v15i3.5278

[28] Prayitno, A., et al. (2021). A systematic review of federated learning in the healthcare domain. Applied Sciences, 11(23), 11191. https://doi.org/10.3390/app112311191

[29] Saba, L., Biswas, M., Kuppili, V., et al. (2019). The present and future of deep learning in radiology. European Journal of Radiology, 114, 14–24. https://doi.org/10.1016/j.ejrad.2019.02.038

[30] Shur, J., Blackledge, M., & Doran, S. J. (2021). Radiomics in oncology: A practical guide. Radiographics, 41(6), 1717–1731. https://doi.org/10.1148/rg.2021210029

[31] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7