International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| 9th ICAILLMBA (ALLAM) 2025-Conference Proceedings
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.ICAILLMBA-107

Original Article

Relation between Bioinformatics and Computational Statistics in Cancer: A Survey

Mr Chandaka Indra Rao[1], Mr Dadi Yaswanth Kumar[2], Mr Avala Chakrapani[3], Mr Galla Venkataswamy[4], Mr Pagadala Srinivasu[5]
[1]Assistant Professor Department of CSE (AI&ML, DS), Avanthi Institute of Engineering & Technology, Cherukupally (P), Bhogapuram (M), Near Tagarapuvalasa, Nh 16, Kotabhogapuram, Andhra Pradesh -535006.
[2]Programmer, Department of CSE (AI&ML, DS), Avanthi Institute of Engineering & Technology, Cherukupally (P), Bhogapuram (M), Near Tagarapuvalasa, Nh 16, Kotabhogapuram, Andhra Pradesh -535006.
[3]Assistant Professor, Department of Computer Science and Engineering, Raghu Engineering College (A), Dakamarri, Visakhapatnam, Andhra Pradesh-531162.
[4,5]Assistant Professor, Department of CSE (Data Science), Raghu Engineering College (A), Dakamarri, Visakhapatnam, Andhra Pradesh, India.

Abstract - *Cancer research has increasingly embraced bioinformatics and computational statistics to interpret large-scale and complex biological data generated by modern high-throughput technologies [14], [15]. Advances in sequencing and profiling platforms have produced extensive genomic, transcriptomic, and proteomic datasets that require sophisticated analytical techniques for meaningful interpretation [3], [16]. This survey explores how statistical modelling, machine learning, and data-driven approaches enable effective knowledge extraction from cancer data. Key applications such as cancer detection, subtype identification, prognosis estimation, and treatment response prediction are reviewed [4], [6]. In addition, this study addresses methodological challenges, including data heterogeneity, scalability, and interpretability, while discussing recent developments that advance precision oncology and personalised cancer therapy [1], [11].*

Keywords - *Bioinformatics, Computational Statistics, Cancer Research, Genomic Data Analysis, Machine Learning*

1. Introduction

Cancer research has increasingly embraced bioinformatics and computational statistics to interpret large-scale and complex biological data generated by modern high-throughput technologies [14], [15]. Rapid advances in next-generation sequencing, microarray platforms, and mass spectrometry have enabled comprehensive profiling of genomic, transcriptomic, and proteomic landscapes across diverse cancer types. While these technologies generate unprecedented volumes of data, their high dimensionality, noise, and biological variability pose significant analytical challenges, necessitating the use of advanced computational and statistical methodologies for meaningful interpretation [3], [16]. Statistical modelling provides a rigorous foundation for cancer data analysis by supporting hypothesis testing, feature selection, and robust inference under uncertainty. Techniques such as regression analysis, Bayesian models, and regularisation methods are widely used to identify cancer-associated genes, mutations, and molecular pathways while controlling false discovery rates. Complementing these approaches, machine learning techniques enable the discovery of complex, non-linear patterns within multi-omics datasets, facilitating accurate cancer detection, subtype classification, prognosis estimation, and prediction of treatment response [4], [6]. Supervised and unsupervised learning models, including support vector machines, neural networks, and clustering algorithms, have demonstrated strong predictive performance across multiple cancer studies.

In addition, data-driven and integrative approaches play a critical role in combining heterogeneous data sources, such as molecular profiles, clinical records, and imaging data, to generate holistic insights into cancer biology. These frameworks support personalised risk assessment and therapy selection, thereby advancing precision oncology [1], [11]. Despite these advances, challenges related to data heterogeneity, scalability, model interpretability, and clinical translation remain significant. This survey reviews recent methodological developments that address these issues and highlights how the integration of bioinformatics, computational statistics, and machine learning continues to drive progress toward personalised and data-driven cancer therapy.

2. Literature Review

2.1. From (2010–2015)

Between 2010 and 2015, cancer research underwent a major transition toward data-centric methodologies supported by bioinformatics and computational statistics [14]. Researchers increasingly applied statistical learning, clustering, and regularisation techniques to manage high-dimensional gene expression data for accurate cancer classification and prediction [4], [8]. This era also witnessed the rise of network-based approaches and large collaborative initiatives such as The Cancer Genome Atlas (TCGA), which enabled comprehensive genomic profiling across diverse cancer types [16]. By the end of this period, integrative multiomics strategies began to emerge, laying the groundwork for more detailed and precise cancer analyses

[17], provided in the table 2.1 as year-wise literature review on Bioinformatics and Computational Statistics in Cancer Research.

Table 1: Year-Wise Literature Review on Bioinformatics and Computational Statistics in Cancer Research (2010–2015)

Year	Authors	Institution(s)	Focus / Contribution
2010	Golub et al.	Broad Institute, MIT	Early use of statistical learning and clustering methods for cancer classification using gene expression data.
2011	Simon et al.	National Cancer Institute (NCI), USA	Statistical methods for high-dimensional genomic data; regularization techniques for cancer prediction.
2012	Ideker & Sharan	University of California, San Diego	Network-based computational models to analyze cancer pathways and gene interactions.
2013	Weinstein et al.	The Cancer Genome Atlas (TCGA) Consortium	Large-scale genomic characterization of multiple cancer types using statistical and bioinformatics pipelines.
2014	Ramazzotti et al.	University of Milan, NYU	Probabilistic and statistical models for cancer progression and mutation analysis.
2015	Yuan et al.	MD Anderson Cancer Center	Integrative statistical frameworks for multi-omics cancer data analysis.

2.1.1. Good Outcomes during the years from 2010 to 2015

Early statistical learning and clustering approaches enabled accurate cancer classification from gene expression data, laying the foundation for data-driven oncology. The introduction of regularisation, network-based models, and probabilistic frameworks improved the analysis of high-dimensional genomic data and revealed key cancer pathways and progression patterns. Large-scale initiatives such as TCGA further enabled integrative multi-omics analysis, leading to deeper biological insights and more reliable identification of cancer-associated molecular features.

2.2. From (2016–2020)

From 2016 to 2020, cancer research experienced accelerated integration of machine learning and deep learning methods to analyse increasingly complex biomedical datasets [6], [18]. Research efforts progressed from traditional machine learning algorithms to advanced deep learning architectures for survival analysis and outcome prediction based on multi-omics data [2], [12]. During this period, considerable emphasis was placed on curating high-quality, machine learning–ready genomic and transcriptomic datasets to improve model robustness, reproducibility, and generalisability [10]. Several comprehensive reviews published during this time synthesised statistical and machine learning techniques, providing structured guidance for effective multi-omics data integration in cancer research [11], [18], provided in the table 2.2 as year-wise literature review on Machine Learning and Multi-Omics Approaches in Cancer Research.

Table 2: Year-Wise Literature Review on Machine Learning and Multi-Omics Approaches in Cancer Research (2016–2020)

Year	Authors	Institution(s)	Focus / Contribution
2016	Kourou et al.	University of Ioannina, Greece	Review of machine learning techniques (SVM, RF, ANN) applied to cancer diagnosis and prognosis.
2017	Chaudhary et al.	Harvard Medical School	Deep learning models for cancer survival prediction using multi-omics data.
2018	Libbrecht & Noble	University of Washington	Machine learning applications in genomics, emphasizing predictive cancer models.
2019	Lim et al.	Genome Institute of Singapore	Transcriptomic datasets and ML-ready cancer compendiums for predictive analytics.
2020	Nicora et al.	University of Milan	Comprehensive review of ML and statistical tools for multi-omics cancer data integration.

2.2.1. Good Outcomes during the years from 2016 to 2020

During this period, machine learning and deep learning methods significantly improved cancer diagnosis, prognosis, and survival prediction using multi-omics data. The availability of high-quality, ML-ready datasets and comprehensive methodological reviews enhanced model robustness, reproducibility, and effective integration of genomic and transcriptomic information, strengthening data-driven cancer research.

2.3. From (2021–2025)

From 2021 to 2025, cancer research has increasingly been shaped by sophisticated computational workflows and AI-enabled methodologies designed to enhance clinical applicability [3], [9]. Research during this period prioritised high-confidence genome sequencing analysis, precise variant annotation, and AI-driven biomarker identification [3], [7]. Deep learning techniques were widely adopted for prognostic modelling using integrated genomic and transcriptomic datasets, while growing attention was given to knowledge-guided and multi-omics frameworks to improve interpretability and advance precision oncology [1], [11], provided in the table 2.3 as Recent Advances.

Table 3: Recent Advances in Computational and AI-Driven Cancer Research (2021–2025)

Year	Authors	Institution(s)	Focus / Contribution
2021	Cortés-Ciriano et al.	EMBL-EBI, Wellcome Sanger Institute	Computational analysis pipelines for cancer genome sequencing and variant interpretation.
2022	Zou et al.	Tsinghua University	Statistical learning and AI-based models for cancer biomarker discovery.
2023	Lee	Seoul National University	Deep learning approaches for cancer prognosis using genomic and transcriptomic data.
2024	Mao et al.	University of Texas, MD Anderson	Knowledge-guided machine learning models addressing interpretability and clinical relevance.
2025	Acharya & Mukhopadhyay	University of Kalyani	Advanced ML and deep learning frameworks for precision oncology using multi-omics data.

2.3.1. Good Outcomes during the years from 2021 to 2025

This period marked the maturation of AI-driven cancer research, with robust computational pipelines enabling accurate genome sequencing analysis and variant interpretation. Knowledge-guided and deep learning models improved biomarker discovery, prognostic accuracy, and clinical interpretability, strengthening the translation of multi-omics insights into precision oncology applications.

3. Methodology Materials and Methods

This study adopts a structured methodology to review statistical models, machine learning techniques, and data-driven approaches used in cancer bioinformatics, with a focus on analysing genomic, transcriptomic, and proteomic data [11], [14].

3.1. Review of Statistical Models

Statistical models play a central role in managing high-dimensional biological data by enabling the identification of significant genes, mutations, and protein expressions while effectively accounting for noise, variability, and uncertainty inherent in cancer omics datasets [14]. High-throughput technologies often generate data with far more features than samples, making conventional analysis unreliable without rigorous statistical control. Techniques such as linear models and regression analysis are widely used to quantify associations between molecular features and cancer phenotypes, while regularisation methods help prevent overfitting in high-dimensional settings. Bayesian frameworks further enhance cancer data analysis by incorporating prior biological knowledge and providing probabilistic estimates that capture uncertainty in model parameters and predictions. In addition, hypothesis testing and multiple testing correction procedures are essential for controlling false discovery rates and ensuring the reliability of identified cancer-related biomarkers [4], [13]. Together, these statistical approaches provide a principled and interpretable foundation for cancer research, supporting robust inference and reproducible results. By ensuring analytical reliability and biological relevance, statistical models remain indispensable for downstream machine learning, integrative analyses, and clinical translation in precision oncology.

Table 4: Role of Statistical Models in High-Dimensional Cancer Omics Data Analysis

Aspect	Description
Data Type	Genomic, transcriptomic, proteomic data
Problem	High dimensionality, noise, variability
Statistical Models Used	Linear models, regression, Bayesian models, hypothesis testing
Output	Significant genes, mutations, protein expressions
Advantage	Handles uncertainty and reduces false discoveries

Table 4 demonstrates the application of statistical models to genomic, transcriptomic, and proteomic datasets to address challenges related to dimensionality, noise, and biological variation. These methods enhance analytical robustness and support reliable identification of cancer-related molecular features [14], [17].

3.2. Review of Machine Learning Techniques

Machine learning techniques play a vital role in cancer research by enabling effective pattern recognition, cancer subtype classification, biomarker discovery, and clinical outcome prediction from complex multi-omics datasets [6], [8]. Supervised learning models, such as support vector machines, random forests, and neural networks, are widely used to predict disease risk, patient survival, and treatment response by learning discriminative patterns from labelled genomic, transcriptomic, and proteomic data. These models are particularly valuable for handling high-dimensional data and capturing complex, non-linear relationships that are difficult to model using traditional statistical approaches [4], [18]. Unsupervised learning methods, including clustering and dimensionality reduction techniques, facilitate the discovery of novel cancer subtypes and molecular signatures by uncovering intrinsic structures within heterogeneous datasets. Such approaches support exploratory analysis and hypothesis generation, revealing previously unknown biological patterns. In addition, ensemble and deep learning models have demonstrated improved predictive performance by integrating diverse omics features and learning hierarchical representations. Despite their strong performance, challenges related to interpretability, data imbalance, and model generalisation remain. Nevertheless, machine learning techniques continue to significantly enhance the analytical capabilities of cancer research, contributing to more accurate diagnosis, prognosis, and personalised treatment strategies in precision oncology.

Table 5: Machine Learning Techniques for Multi-Omics Cancer Data Analysis

Category	Description
Data Type	Genomic, transcriptomic, proteomic (multi-omics) data
ML Models Used	Supervised (SVM, Random Forest, Neural Networks), Unsupervised (K-means, Hierarchical clustering)
Key Tasks	Pattern recognition, cancer subtype classification, biomarker discovery, outcome prediction
Input Features	Gene expression levels, mutation profiles, protein abundance
Output	Cancer classes, risk scores, significant biomarkers
Advantage	Handles complex, non-linear relationships in high-dimensional data

Table 5outlines the use of machine learning models across genomic, transcriptomic, and proteomic data for tasks such as subtype classification and outcome prediction. These techniques support robust biomarker identification and enhance predictive performance in cancer research [2], [12].

3.3. Review of Data-Driven Approaches

Data-driven approaches play a crucial role in modern cancer research by integrating heterogeneous data sources, including genomic, transcriptomic, proteomic, clinical, and imaging data, to uncover latent relationships and molecular interactions underlying disease progression [11], [19]. Unlike single-modality analyses, these approaches leverage data fusion and feature-level integration techniques to combine complementary information across diverse datasets, enabling a more comprehensive understanding of cancer biology. Network-based and systems-level models further enhance this integration by capturing complex interactions among genes, proteins, and pathways, which are often disrupted in cancer. By analysing multi-source data collectively, data-driven frameworks improve cancer detection accuracy, prognostic assessment, and treatment response prediction. These methods support the development of personalised risk profiles and therapy recommendations tailored to individual patients, thereby advancing precision oncology. Additionally, data-driven approaches facilitate the identification of clinically relevant biomarkers and molecular signatures that may not be detectable through isolated analyses. Despite challenges related to data standardisation, scalability, and interpretability, continued advances in computational infrastructure and integrative modelling techniques are addressing these limitations. Overall, data-driven approaches provide a powerful foundation for translating complex cancer data into actionable clinical insights and personalised treatment strategies.

Table 6: Data-Driven Integration Approaches for Personalised Cancer Analysis

Aspect	Description
Data Sources	Genomic, transcriptomic, proteomic, clinical, imaging data
Integration Methods	Data fusion, feature-level integration, network-based models
Analysis Goal	Discover hidden molecular relationships and interactions
Applications	Cancer detection, prognosis, treatment response prediction
Outcome	Personalized risk scores and therapy recommendations
Advantage	Provides a holistic view of cancer biology

Table 6 highlights data fusion, feature-level integration, and network-based models that combine genomic, transcriptomic, proteomic, and clinical information. These approaches enable personalised risk assessment and therapy recommendations, advancing precision oncology [9], [11].

4. Results and Discussion

4.1. Results

Statistical models applied to genomic, transcriptomic, and proteomic datasets have effectively addressed challenges associated with high dimensionality, noise, and biological variability commonly encountered in cancer research [14]. Approaches such as linear and regression models, Bayesian methods, and hypothesis testing frameworks enabled the reliable identification of cancer-related genes, mutations, and molecular features while maintaining interpretability and statistical rigor [13], [17]. These methods provided a strong analytical foundation by controlling uncertainty and reducing false discoveries.

In parallel, machine learning techniques demonstrated strong performance in pattern recognition, cancer subtype classification, and clinical outcome prediction using multi-omics data [6]. Supervised learning models achieved high predictive accuracy in risk stratification and survival analysis, whereas unsupervised methods uncovered intrinsic data structures and revealed molecular heterogeneity within cancer populations [4], [18]. Such capabilities are essential for understanding disease complexity and supporting personalised medicine. Furthermore, data-driven approaches enabled comprehensive integration of heterogeneous datasets, including molecular, clinical, and imaging information, leading to the discovery of hidden molecular interactions [11], [19]. This integrative analysis significantly improved prognostic accuracy and treatment response prediction, reinforcing the importance of combining statistical, machine learning, and data-driven methodologies to advance precision oncology.

4.2. Discussion

Overall, statistical models, machine learning techniques, and data-driven approaches work synergistically to advance cancer data analysis by addressing complementary analytical challenges across diverse biomedical datasets [11]. Statistical models form the backbone of analytical rigor by ensuring robustness, interpretability, and reliable inference, particularly in high-dimensional and noisy cancer omics data. Their ability to control uncertainty and false discovery rates supports trustworthy identification of biologically meaningful features. In contrast, machine learning techniques excel at capturing complex, non-linear relationships that are difficult to model using traditional statistical methods. By leveraging supervised and unsupervised learning, these techniques enhance cancer detection, subtype classification, and outcome prediction, enabling more accurate and scalable analytical solutions. Data-driven frameworks further strengthen cancer analysis by integrating heterogeneous data sources, including genomic, transcriptomic, proteomic, clinical, and imaging data, to provide holistic insights into cancer biology. This integrative perspective supports comprehensive disease characterisation and personalised risk assessment. Together, the combined use of statistical, machine learning, and data-driven approaches significantly improves cancer diagnosis, prognosis, and personalised treatment strategies, thereby advancing the goals of precision oncology and data-driven clinical decision-making [1], [9].

5. Conclusion

This survey comprehensively reviewed the role of bioinformatics and computational statistics in modern cancer research, with an emphasis on genomic, transcriptomic, and proteomic data analysis [14], [15]. The literature review highlights a clear evolution from traditional statistical methods to advanced machine learning, deep learning, and data-driven multi-omics integration approaches [6], [11]. The findings demonstrate that statistical models ensure robustness and interpretability, machine learning techniques enable accurate prediction and subtype classification, and data-driven approaches support holistic and personalised cancer analysis. Overall, the integration of these computational methodologies plays a critical role in advancing precision oncology and improving clinical decision-making in cancer care [1], [9].

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest concerning the publishing of this paper

References

[1] D. Acharya and A. Mukhopadhyay, “Machine learning frameworks for multi-omics data integration in precision oncology,” *Briefings in Functional Genomics*, vol. 24, no. 2, pp. 123–138, 2025.

[2] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, “Deep learning–based multi-omics integration robustly predicts survival in liver cancer,” *Clinical Cancer Research*, vol. 24, no. 6, pp. 1248–1259, 2018, doi: 10.1158/1078-0432.CCR-17-0853.

[3] I. Cortés-Ciriano, D. C. Gulhan, J. J. K. Lee, G. E. M. Melloni, P. J. Park, and G. Getz, “Computational analysis of cancer genome sequencing data,” *Nature Reviews Genetics*, vol. 23, no. 5, pp. 298–314, 2022, doi: 10.1038/s41576-021-00431-y.

[4] T. R. Golub *et al.*, “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, no. 5439, pp. 531–537, 1999, doi: 10.1126/science.286.5439.531.

[5] T. Ideker and R. Sharan, “Protein networks in disease,” *Genome Research*, vol. 18, no. 4, pp. 644–652, 2008, doi: 10.1101/gr.071852.107.

[6] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[7] M. Lee, “Deep learning techniques with genomic data in cancer prognosis,” *Biology*, vol. 12, no. 7, p. 893, 2023, doi: 10.3390/biology12070893.

[8] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015, doi: 10.1038/nrg3920.

[9] S. B. Lim, S. J. Tan, W. T. Lim, and C. T. Lim, “An extracellular vesicle–based liquid biopsy for lung cancer,” *Scientific Data*, vol. 6, p. 246, 2019, doi: 10.1038/s41597-019-0207-2.

[10] R. Simon, E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao, *Design and Analysis of DNA Microarray Investigations*. New York, NY, USA: Springer, 2011.

[11] G. Nicora, F. Vitali, A. Dagliati, N. Geifman, and R. Bellazzi, "Integrated multi-omics analyses in oncology: A review of machine learning methods and tools," *Frontiers in Oncology*, vol. 10, p. 1030, 2020, doi: 10.3389/fonc.2020.01030.

[12] D. Ramazzotti *et al.*, "CAPRI: Efficient inference of cancer progression models from cross-sectional data," *Bioinformatics*, vol. 31, no. 18, pp. 3016–3026, 2015, doi: 10.1093/bioinformatics/btv296.

[13] R. Simon and R. M. Subramanian, "Statistical aspects of genomic data analysis in cancer research," *Journal of Biopharmaceutical Statistics*, vol. 21, no. 3, pp. 412–429, 2011.

[14] J. N. Weinstein *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013, doi: 10.1038/ng.2764.

[15] Y. Yuan *et al.*, "Assessing the clinical utility of cancer genomic and proteomic data across tumor types," *Nature Biotechnology*, vol. 32, no. 7, pp. 644–652, 2014, doi: 10.1038/nbt.2940.

[16] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature Genetics*, vol. 51, no. 1, pp. 12–18, 2019, doi: 10.1038/s41588-018-0295-5.

[17] T. Ideker, N. J. Krogan, and R. Sharan, "Network-based approaches to cancer genomics," *Genome Biology*, vol. 12, no. 2, p. 205, 2011.

[18] L. Mao, H. Wang, L. S. Hu, and N. L. Tran, "Knowledge-informed machine learning for cancer diagnosis and prognosis," *Briefings in Bioinformatics*, vol. 25, no. 1, p. bbad501, 2024, doi: 10.1093/bib/bbad501.

[19] Y. Yuan and G. Getz, "Data-driven integration of multi-omics data for precision oncology," *Annual Review of Biomedical Data Science*, vol. 2, pp. 203–227, 2019.