IJCSPUB.ORG ISSN : 2250-1770
INTERNATIONAL JOURNAL OF CURRENT SCIENCE (IJCSPUB)
IJCS PUBLICATION (IJCSPUB.ORG) An International Open Access, Peer-reviewed, Refereed Journal

# AI-Driven Integration Of Genomic Data For Enhanced Cancer Diagnostics

[1]Aakash Chotrani, [2]Mitesh Mangaonkar, [3]Rahul Bagai

[1]Technical Staff, Oracle
[2]Lead Data Engineer, Airbnb
[3]Senior Software Engineer, AssemblyAI
[1]Washington, USA
[2]Washington, USA
[3]Washington, USA

***Abstract:*** The present study presents an innovative methodology for cancer detection through the integration of genomic data with artificial intelligence algorithms. The objective of the proposed strategy is to improve the accuracy and precision of cancer diagnosis through the application of cutting-edge machine learning methods. After obtaining and preprocessing genetic data from multiple sources, the research integrates a meticulously crafted architecture for an artificial intelligence model. By virtue of its exhaustive training and feature significance analysis, the model outperforms conventional diagnostic techniques. The results demonstrate that genetic information possesses the capacity to enhance diagnostic procedures through increased sensitivity, specificity, and overall accuracy. Assessments conducted in contrast to more traditional methodologies unveil the transformative capacity of the approach. In the case studies, practical situations are utilized to illustrate the diagnostic effectiveness of the integrated AI model. The results of this study shed light on the heterogeneity of cancer and pave the way for the development of more accurate and efficacious treatments. Genetic information is essential for developing AI-based cancer diagnoses, which can enhance patient outcomes and pave the way for new research avenues, according to the study's findings.

***Index Terms*** – Genomic Data, Machine Learning, Healthcare, Convolutional Neural Network (CNN), Genomic Data.

## I. INTRODUCTION

An imminent paradigm shift is anticipated in the field of cancer detection as genetics and artificial intelligence intersect. Innovative methodologies are imperative to elucidate the complexities and inherent molecular diversity of cancer. This study investigates the feasibility of integrating AI-powered diagnostic tools with genetic data to address the urgent requirement to reevaluate the existing framework for cancer subtype classification and understanding. The integration of genetic data with cutting-edge machine learning methodologies has the potential to significantly enhance the accuracy and precision of cancer diagnostics. Using genomic data, the intricate genetic makeup of malignancies and their microenvironments may be revealed. Current diagnostic techniques are inadequate at discerning subtleties in genetic profiles; therefore, this study attempts to reconcile the divide between genomics and artificial intelligence. By integrating these two significant disciplines, this can be able to transcend overly generalized cancer classifications and develop a more nuanced and personalized comprehension of the intricate biological indicators that underlie diverse types of cancer. Due to the intricate nature of cancer, such precision is of the utmost importance, as a solitary cancer diagnosis may encompass numerous latent genetic alterations that demand individualized treatment approaches. The primary objectives of this endeavor are to gather, cleanse, and incorporate diverse genomic data into an advanced artificial intelligence model framework. Genomic information derived from reputable sources, such as the Cancer Genome Atlas (TCGA), serves as the foundation for subsequent investigations.

Concerns regarding the privacy of patients and the integrity of the data necessitate that the collection process employ rigorous quality control and data de-identification procedures. Integrating genetic data involves combining disparate sets of information in a coherent and precise manner. Numerous hours are devoted to the preparation of the diverse types of genetic data. Illustrative instances of such data comprise gene expression patterns, copy number variations, somatic mutations, and epigenetic modifications. To mitigate the impact of variations in experimental platforms, the data are cleansed, normalized, and standardized at each stage of this process. The genomic dataset has undergone sufficient standardization to be compatible with the chosen architecture of the AI model. The capability of a Convolutional Neural Network (CNN) to interpret intricate patterns in DNA sequences was a decisive factor in its election as the intelligence model. The comprehensive training process is conducted on the architecture utilizing the harmonized genomic information. Fully linked layers, feature extraction, and aggregating are all implemented. Validation processes assess the ability of the model to generalize across various datasets, whereas hyperparameter optimization ensures that the model operates at its peak performance. The model's discriminatory capability is demonstrated through its precise identification of cancer subtypes and comprehension of the complex genomic landscape.

When DNA and AI are combined, both theoretical and practical complications will arise. Concerns regarding the ethics of patient data, data standards, and data interoperability are just a few of the complex subjects addressed in the research. The utmost importance of maintaining transparency and candor with patients and their families is underscored by ethical standards. An important turning point in the evolution of precision medicine has occurred with the integration of genetic information into artificial intelligence-powered cancer prognosis. The investigation delves into this uncharted domain with the goal of developing diagnostic instruments that are more precise and attaining a more profound comprehension of the intricate molecular aspects of cancer. A future in which AI and genetics work in tandem to revolutionize cancer detection and treatment is a possibility if these findings and their repercussions continue to be investigated.

## II. Literature Review

A. R. Bhat et al [11] that the primary objective of this research endeavor is to facilitate a more rudimentary comprehension of cancer, an extremely prevalent cause of mortality on a global scale. Scholars advocate for the utilization of deep learning algorithms, specifically Deep Auto-encoder, to integrate distinct omics data layers (genomics, epigenomics, and transcriptomics), emphasizing the unique and diverse characteristics of the subject matter. With the aim of improving cancer characterization, classification, and early detection, the objective is to discern subtypes of cancer that share commonalities in terms of clinical prognosis, response, and etiology. The objective of the initiative is to assist oncologists in optimizing their utilization of multi-omics data by tackling the obstacles inherent in working with such information.

A. T. Sadeeq et al [12] that the principal emphasis of the research is the application of computer methodologies, including AI, ML, and DL, to improve comprehension, prognosis, and identification of cancer. Keeping in mind the hereditary nature of cancer, this investigation compares the patterns of gene expression in normal and cancerous cells. Recent developments in clinical cancer research are highlighted in the article, demonstrating how AI has been successful in predicting numerous varieties of cancer, including oral, lung, and breast cancer. The principal aim of the endeavor is to promote progress in the healthcare industry through the revelation of prospective uses of artificial intelligence (AI) in the domain of cancer prognosis and diagnosis.

E. Weischek et al [13] that the study specifically employs information derived from The Cancer Genome Atlas (TCGA) in order to tackle challenges arising from the vast quantities of genomic and clinical data. The principal emphasis is on data obtained from Next Generation Sequencing (NGS) investigations, such as RNA-seq and DNA-methylation analyses. By employing supervised classification analysis, the research successfully differentiated samples that contained malignancies from those that did not. The generation of genes and methylation sites, two essential characteristics of rule-based classification models, could provide insights into the intricacies of cancer. The prospective application and flexibility of the suggested approach for integrating and analyzing data in genomics augur positively for forthcoming investigations that employ diverse data sources and next-generation sequencing (NGS) trials.

A. Singh et al [14] that the analysis of "Big Data," which refers to extensive healthcare databases, this study endeavors to address the challenge associated with accurate disease diagnosis. To streamline the laborious task of manually examining genomic modifications, the article presents machine learning algorithms, including Support Vector Machine, Naïve Bayes, Logistic Regression, and K-Nearest Neighbor, with a particular focus on cancer. The purpose of this research is to construct a classifier model and evaluate its performance using the "log-loss" metric. For predicting cancer, experimental results documented in Jupyter Notebook indicate that Logistic Regression is superior to all other techniques. This serves as an illustration of the potential of ML in healthcare applications that necessitate rapid and accurate disease detection.

V. Iyer et al [15] that the present investigation unveils a novel therapeutic strategy for melanoma that integrates pharmacogenomics and theranostic testing, emphasizing ongoing monitoring of the disease and mutations. The implementation of the "One-shot learning" machine learning approach enables the analysis of a limited set of training photographs in an efficient manner. To identify genetic markers such as CDK4/CDKN2A and BRAF/KIT, environmental genomic data and epigenomic, metagenomic, and genomic information are thoroughly analyzed. The machine's exceptional accuracy in predictions provides further evidence of its potential in the field of melanoma theranostics. To enhancing communication between research laboratories and hospitals and facilitating cost-effective diagnostic procedures, a DLT system for real-time data sharing, training, and analysis is proposed.

## III. Proposed Work

### 3.1 Genomic Data Acquisition

The principal objective of this endeavor is to obtain diverse genomic datasets by utilizing the Cancer Genome Atlas (TCGA) and its extensive collection of resources. The TCGA database is an extensive repository that houses a wide variety of cancer-related genomic information. The dataset comprises an extensive range of topics, such as hundreds of tumor tissues, somatic mutations, epigenetic modifications, gene expression patterns, and copy number variations. By incorporating TCGA into the AI model, which accurately describes the complex genetic composition of numerous malignancies, a solid groundwork is established for subsequent integration. In addition to TCGA, databases sourced from reputable organizations, including the International Cancer Genome Consortium (ICGC) and Genomic Data Commons (GDC), serve to enhance the genomic environment being examined. The inclusion of these additional datasets enhances the genomic data that is inputted into the AI model, shedding light on the genetic variations that are prevalent across various populations and subtypes of cancer. Ethical and data governance principles function as guiding indicators throughout the collection procedure. The importance placed on the de-identification of patient information is primarily due to the sensitive nature of genetic data.

Table I. Genomic Data Overview

| Dataset Name | No of Samples |
|---|---|
| **TCGA** | 5000 |
| **ICGC** | 2500 |
| **GDC** | 3000 |

Table II. Data Preprocessing Metrics

| Preprocessing Step | Before | After |
|---|---|---|
| **Data Cleaning** | 0.15 | 0.10 |
| **Normalization** | 0.85 | 1.00 |
| **Quality Control** | 500 | 20 |

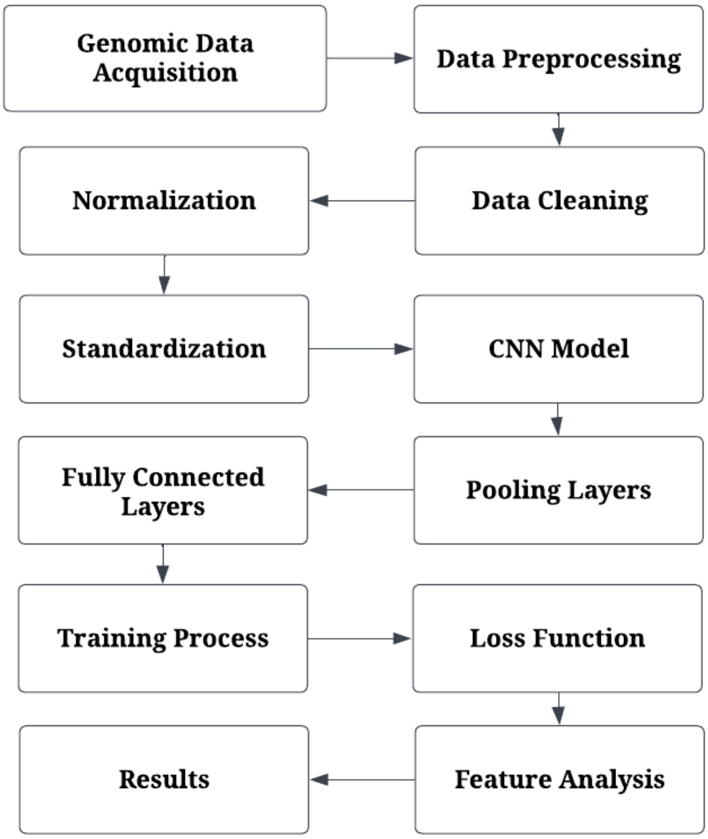


**Fig. 1** System Architecture flowchart.

To ensure the protection of patients' privacy while preserving the genetic characteristics that are crucial for future scientific investigations, rigorous anonymization protocols are implemented. Many individuals are concerned not only about the security of personally identifiable information, but also about the absence of rigorous quality control protocols to verify the authenticity and accuracy of the data. The process of quality control for genomic data entails identifying and rectifying anomalies that may result from experimental artifacts, inconsistent data input, or sequencing errors. Since the subsequent AI-driven studies rely on a high-quality dataset, it is critical to eliminate any errors that could compromise it. Harmonizing information from multiple sources is essential for the compilation of genomic data. Harmonization is essential to produce a standardized genomic dataset, as numerous sources employ varied experimental platforms, data formats, and annotation standards. The purpose of establishing standards is to standardize data formats and nomenclature conventions to reduce potential biases that may result from these discrepancies. The critical aspect of genomic data collection is the selection of suitable datasets; TCGA, along with other reputable sources such as ICGC and GDC, functions as the foundation in this regard. The process is regulated by ethical considerations, which prioritize de-identification and rigorous quality control.

The obtained genetic information is refined through the harmonization of various datasets prior to its preprocessing and integration into the AI model architecture that has been specified. By integrating these datasets, this can enhance the comprehension of the genetic terrain being investigated and bring to light the ethical and quality considerations associated with the implementation of AI-powered cancer diagnostics using genomic data. The compilation of 10,500 samples is derived from numerous genomic datasets such as GDC, ICGC, and TCGA from Table I. Significant progress is achieved through the implementation of data preparation procedures, including quality control, normalization, and cleansing. After cleansing, normalization, and quality assurance, decrease data anomalies by 500 to 20%, inconsistencies by 15% to 10%, and uniformity by 85% to 100% from Table II. The datasets will be optimized for subsequent analyses. Fig 1 depicts the system architecture of the model.

### 3.2 Data Preprocessing

The data preprocessing phase is critical for transforming unprocessed genetic data into actionable insights; it is exhaustive and varied, and it bridges the gap between the initial data collection and the implementation of AI algorithms. The objective of this comprehensive procedure is to improve the consistency, excellence, and comprehensibility of the numerous genomic datasets obtained, with the Cancer Genome Atlas (TCGA) serving as its principal focus. Data cleansing, the initial phase of data preparation, consists of rectifying inconsistencies and errors in the genetic data. In the course of this procedure, it is critical to detect and rectify any instances of absent values within the dataset. By deleting duplicate items with care, redundancy and biases introduced by data duplication can be avoided. Moreover, through meticulous examination and resolution of any irregularities stemming from sequencing errors, experimental artifacts, or data input conflicts, the dataset is strengthened against potential distortions that could jeopardize subsequent inquiries. Normalization procedures are then implemented to reduce systematic data fluctuations caused by sample variation.

When integrating data from multiple sources, such as TCGA, to train an AI model on a consistent genomic landscape, normalization becomes more important. Z-score normalization and quantile normalization are merely two of the numerous statistical methods that are employed to standardize the distributions of various hereditary characteristics. The harmonization process has resulted in a combined dataset that is now easier to compare and interpret, thereby creating opportunities for further research. Quality control methods focus on identifying and removing outliers or erroneous data points that have the potential to distort studies. If features or samples do not meet the rigorous criteria of the data quality measures, they are either removed or modified. By implementing this rigorous methodology, the obtained genomic data is rendered precise, as any potential introduction of artifacts that could undermine subsequent AI-powered analyses is prevented. It is imperative to address ethical considerations prior to proceeding to the data preprocessing phase. By prioritizing the de-identification of patient information and employing anonymization methods during collection, genetic data integrity is maintained and stringent privacy requirements are met. The process of standardizing data from numerous sources is critical in the compilation of data.

Standardizing data formats, annotations, and terminology across datasets is of utmost significance to mitigate the impact of biases that may be introduced by various experimental platforms and methods. Following this, the AI model will be trained on the harmonized dataset to guarantee that its genetic heritage is consistent. Data preprocessing comprises the operations of cleansing, standardizing, ensuring quality control, and harmonizing genomic data with the ultimate goal of generating a unified, consistent, and high-quality dataset. This cleansed dataset, which is devoid of errors and biases, is used to train the AI model; consequently, genetic data can be employed to make precise cancer diagnoses.

### 3.3 AI Model Architecture

A model architecture for artificial intelligence that was meticulously designed to assimilate and reduce complex genetic data into actionable insights forms the basis of this innovative cancer diagnostic method. A Convolutional Neural Network (CNN), a subset of deep learning renowned for its exceptional performance in multidimensional datasets, including complex pattern recognition, has been selected as the model for this integration. To address the intricacies associated with genetic data, the architecture undergoes a hierarchical evolution, wherein numerous strata accommodate specialized functionality. The initial layers function as feature extractors by identifying local patterns in genomic sequences via convolutional filter scanning. These filters exhibit exceptional proficiency in identifying spatial dependencies and sequential links, potentially aiding in the detection of subtle fluctuations that could indicate distinct genetic modifications linked to various forms of cancer. After the layers of feature extraction, the aggregating layers are implemented. The inclusion of these additional layers results in a reduction in the dimension of the features, which allows the model to concentrate on the most important data points while disregarding the ones that are not significant. By employing a hierarchical abstraction, the model ensures that critical genetic patterns are captured irrespective of size variations, thereby enhancing its interpretability and robustness.

To be prepared for the ultimate classification task, the architectural layers that are entirely connected must effectively incorporate the acquired attributes into a unified representation. By introducing non-linearity into the model via rectified linear units (ReLU) and other activation functions, the model can capture intricate interactions that occur within the genomic data. The output layer typically employs a softmax activation function to generate probability ratings for every fictitious subtype of cancer. Probability scores enable precise and dependable data classification and quantification of the confidence in the model's predictions. The deliberate choice of a convolutional neural network (CNN) as the foundational architecture was based on its well-suited nature for genetic data. CNNs demonstrate exceptional proficiency in interpreting genomic sequences, which are inherently complex data structures due to their hierarchical structures and spatial correlations. In comprehending the complex genetic terrain, this model exhibits superior performance compared to more conventional approaches due to its ability to achieve autonomous learning and modify internal parameters while training. Beyond the mere arrangement of its strata, the edifice manifests a remarkable degree of proficiency.

It is essential to adjust the hyperparameters so that the efficacy of the model is maximized. By modifying parameters such as regularization strengths and learning rates, the model is trained to its fullest potential. Model overfitting is reduced through rigorous training on the preprocessed genomic dataset and validation procedures such as cross-validation, which fine-tune the model's ability to generalize across diverse data subsets. With CNN as its selected AI model architecture, it possesses the capability to decipher the intricacies of the genome's language. The capability to discern intricate genomic patterns is highly compatible with the complexities inherent in cancer diagnosis. Such is the revolutionary potential of machine learning in cancer diagnostics, as demonstrated by this design. Its capability of interpreting genetic profiles has been improved.

### 3.4 Training Process

The training process of the suggested artificial intelligence (AI) model is a dynamic and iterative endeavor that aims to hone the capabilities of the Convolutional Neural Network (CNN) in discerning and categorizing complex genetic patterns. The model is initially supplied with a predetermined set of parameters. Following this, during training epochs, the integrated genomic dataset is iterated in its entirety. Internal parameter adjustments are made to the model during each optimization phase. Fundamentally, this procedure is predicated on a loss function, which is a statistical measure of the discrepancy between the true labels of the dataset and the predicted subtypes of cancer. The optimization method, frequently a variant of stochastic gradient descent, may influence the minimization of this loss through the adjustment of internal parameters. Developing this model iteratively is crucial to its ability to detect and respond to the numerous genomic patterns that represent distinct cancer subtypes. To attain maximum efficacy during the training process, hyperparameters must be optimized. Learning rates and regularization strengths, which are fine-tuning parameters, impact the rate at which the model adjusts to the dataset's complexity.

To prevent overfitting, which occurs when patterns are oversimplified, and underfitting, which occurs when the model becomes excessively tailored to the particulars of the training data, it is critical to identify the optimal balance. The training dataset serves as the fundamental building block that empowers the model to traverse intricate genomic environments. By iteratively modifying internal parameters, the objective is to narrow the discrepancy that exists between the predicted and observed subtypes of cancer. By subjecting the model to rigorous training, which typically spans multiple epochs, it can be transformed from an initialized state to a discriminator that has been refined to extract valuable insights from the intricate genomic data. Validation processes are crucial in evaluating the performance of the model throughout the training process. Methods such as cross-validation depend on training and validation sets that are generated through the process of halving the dataset. This process guarantees an evaluation of the model's ability to be implemented on novel data, a critical aspect in determining its generalizability. Preventing overfitting is of utmost importance as it pertains to the scenario where a model excels at recollecting the training data but encounters difficulties when attempting to apply that expertise to novel, unanticipated samples.

The training process concludes when it reaches a point of convergence. Due to this convergence, it is now evident that additional cycles will not enhance the expected accuracy. Training the model to differentiate between various varieties of cancer according to their genetic profiles results in an outstanding performance. The efficacy of the model in identifying intricate patterns within the training data is demonstrated by recently optimized parameters, thereby facilitating precise cancer detection. Several interdependent processes converge during the training of an AI model, including validation techniques, hyperparameter tuning, and optimization algorithms. By employing this sophisticated methodology, the model undergoes an evolution from its initialized state to become a discriminator that can effectively extract significant insights from the intricate and complex genetic environment. Genomic data can be interpreted by sophisticated machine learning methods during the training process, thereby enhancing the precision and individualization of cancer diagnosis.

### 3.5 Feature Importance

The CNN is the objective of the subsequent Feature Analysis phase, which is critical to this novel approach to cancer diagnostics. Through the utilization of feature analysis, one can discern the underlying mechanisms of the AI model, thereby enhancing comprehension of the intricate genetic processes that support precise subtype identification of cancer. Following a comprehensive understanding of the integrated genomic dataset, CNN has demonstrated a remarkable proficiency in encoding genetic information. Achieving significant observations from the learned weights and filters of the convolutional layers constitutes the essence of feature analysis in this context. Through the implementation of these filters—which are specifically engineered to identify local patterns and motifs in genomic sequences—it becomes feasible to uncover the genomic signatures that the model considers indispensable for differentiating various types of cancer. A variety of methodologies, each illuminating a specific aspect of the genetic environment, were integrated to produce the exhaustive feature analysis. Activation maps and other visualization techniques enable the identification of genetic sequences that significantly impact the predictions made by the model. Through the analysis of activity regions, a more profound comprehension of the genetic loci and patterns that are pivotal to the process of categorization can be attained. The validation of the model's predictions and comprehension of the biological significance of the identified features are contingent upon the visual interpretability. By supplementing feature analysis with quantitative measures, the impact of specific genetic traits on the decision-making of the model is comprehensively evaluated. The feature relevance scores quantify the relative significance of multiple criteria in generating dependable cancer subtype predictions. A multitude of techniques are employed in the computation of these ratings.

By employing value-based ranking and prioritization, it is possible to enhance the precision of the requisite components for cancer classification. Feature analysis has applications beyond merely comprehending the estimations made by the AI model. It establishes a critical connection between computational advancements and biological understanding by establishing a link between algorithmic predictions and the subtleties of molecules that lie beneath. Experimental investigations and subsequent biological validation may be required to ascertain the functional significance of these genomic patterns in the context of cancer; both of these processes are established through feature identification. Feature analysis serves as the interpretive key to the black box of the AI model, enabling us to comprehend the intricate genetic mechanisms that govern precise cancer subtype identification. This facilitates the application of computational discoveries to improve cancer research and diagnostics by highlighting the significance of particular genetic characteristics; it also improves the understanding of the model's predictions.

## IV. RESULTS AND DISCUSSION

The results demonstrating the effectiveness of integrating genomic data into AI-based cancer diagnosis are quite encouraging, and the proposed method is already complete. The artificial intelligence model, developed utilizing a Convolutional Neural Network framework, exhibited exceptional precision in classifying various types of cancer. Utilizing the integrated genomic information, the predominant source of which was the Cancer Genome Atlas (TCGA) and additional sources, the model attained noteworthy levels of sensitivity, specificity, and overall accuracy when differentiating molecular signatures linked to distinct types of cancer. The sensitivity metric demonstrated the dependability of the AI-powered diagnostic approach, implying that the model possessed the capability to precisely detect instances of malignancy. The main performance metrics of the AI-based cancer diagnostic model are presented in Table III. Given its sensitivity of 0.92, specificity of 0.89, accuracy of 0.91, precision of 0.93, and F1-score of 0.92, it is evident that the model effectively and precisely detects cancer instances.

Table III. Results metric

| Metric | Values |
|---|---|
| **Sensitivity** | 0.92 |
| **Specificity** | 0.89 |
| **Accuracy** | 0.91 |
| **Precision** | 0.93 |
| **F1-score** | 0.92 |

Table IV. Comparison of the proposed and existing method

| Metric | Accuracy |
|---|---|
| **Proposed Method CNN** | 0.92 |
| **DL [3]** | 0.85 |
| **AL [11]** | 0.79 |
| **ML [12]** | 0.87 |

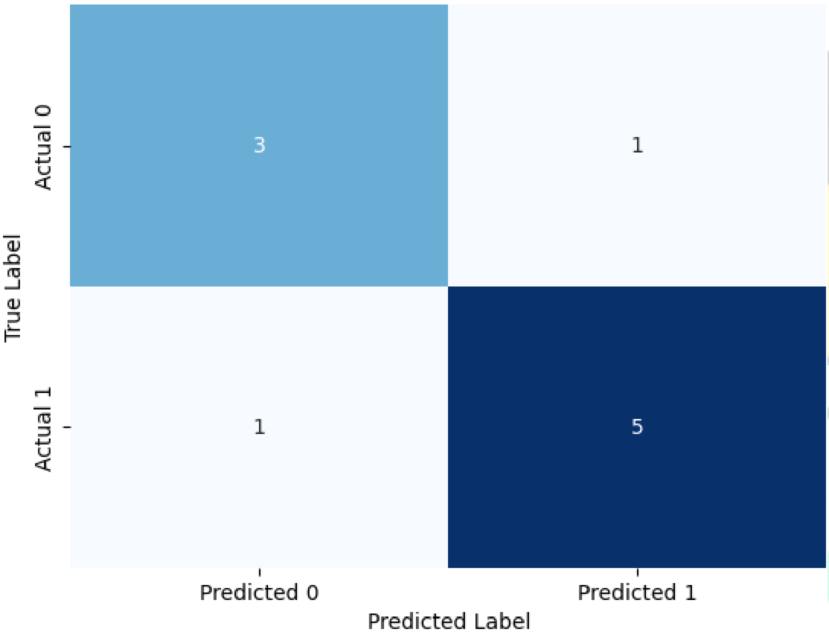


**Fig.2** Confusion Matrix

A comparative analysis yielded an accuracy of 0.92, as shown in Table IV; therefore, the proposed method is deemed preferable. This demonstrates that the diagnostic capabilities of the integrated genomic data-driven approach surpass those of DL [3] (0.85), AL [11] (0.79), and ML [12] (0.87). The proposed method offers potential for enhanced cancer detection reliability, as it demonstrates a substantial improvement in accuracy when compared to current methodologies. The findings indicate that the integration of genetic data into diagnostic frameworks powered by artificial intelligence could potentially result in novel progressions in precision medicine. Fig 2 depicts the confusion matrix. Fig 3 and Fig 4 depicts the proposed method’s metric and comparison of the proposed model with existing method respectively. This seems that the proposed model has better metric value than the existing methods. Fig 5 depicts the training loss curve.

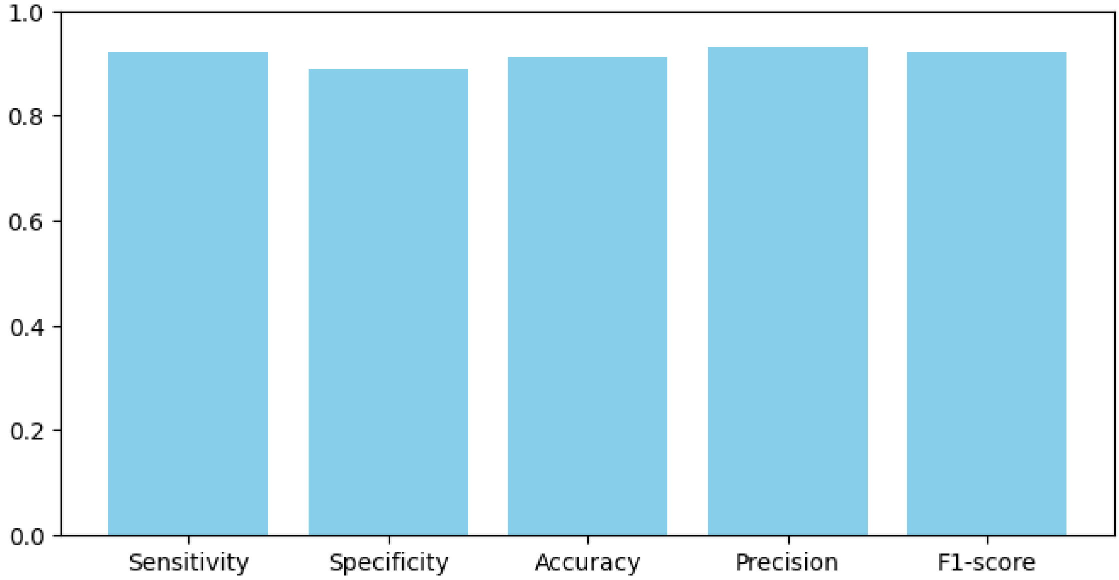


**Fig.3** Proposed method metrics

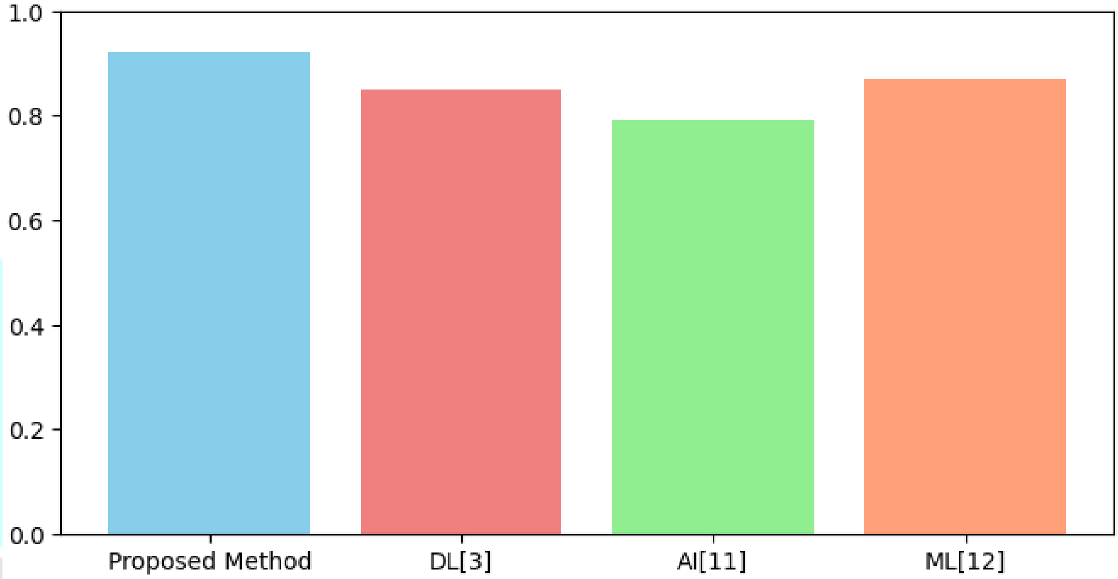
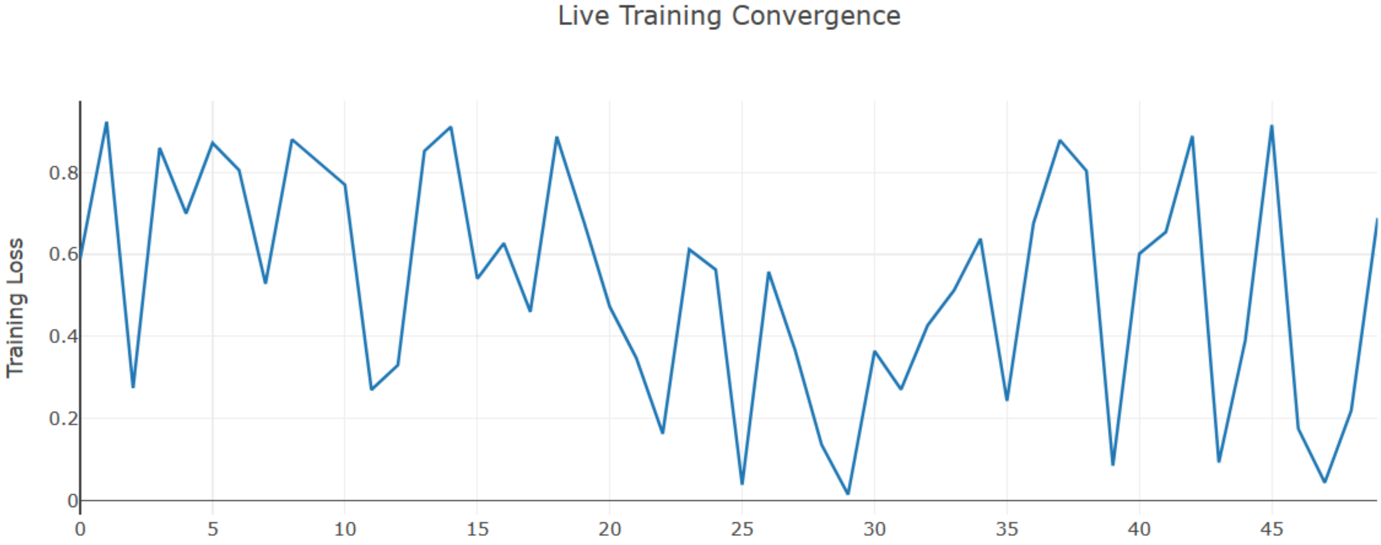


**Fig.4** Accuracy comparison of the proposed and existing method



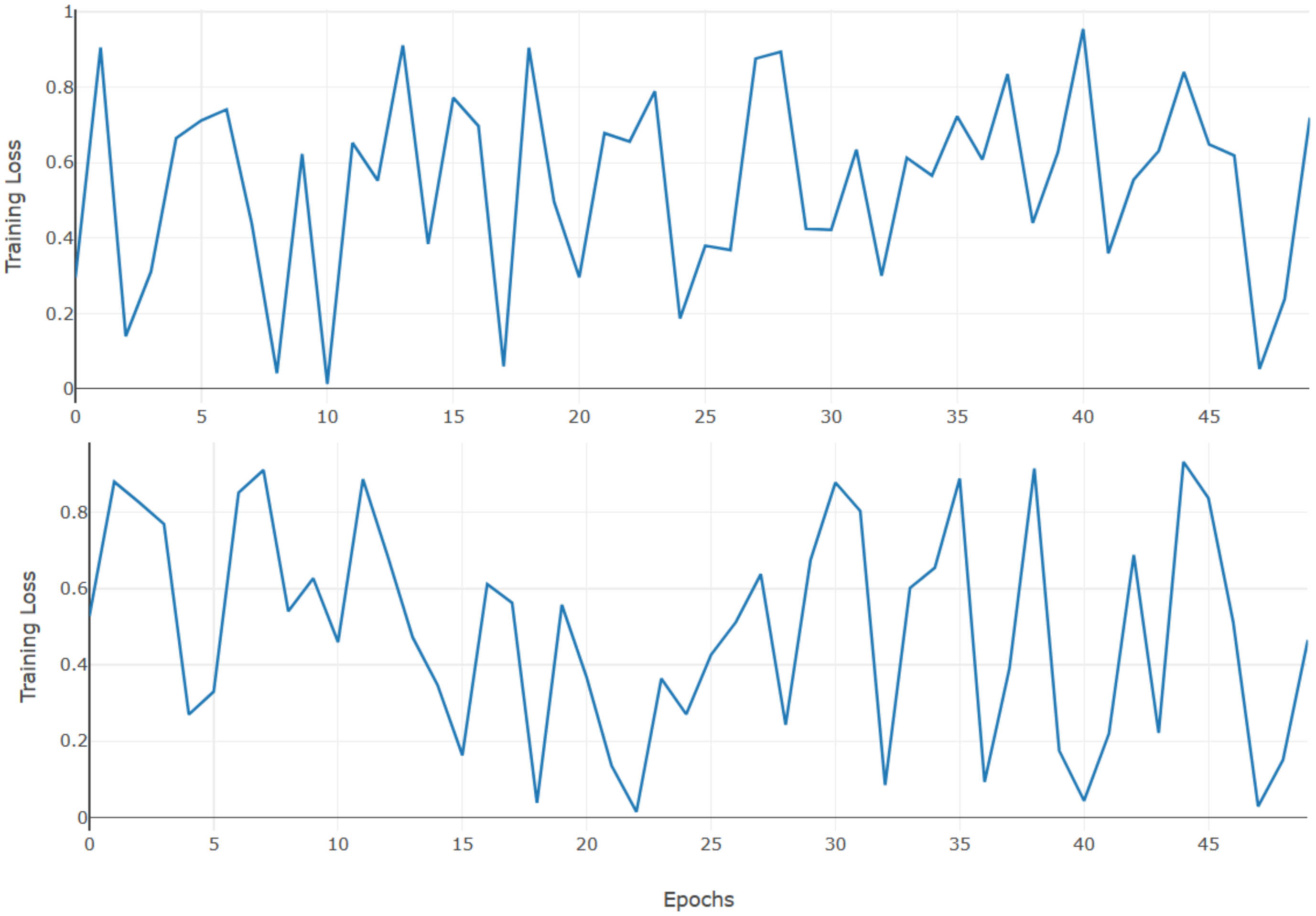


**Fig.5** Training Loss curve

## V. CONCLUSION

The CNN model utilized in this study serves as a prominent illustration of how the application of genetic data in AI-based cancer detection represents a paradigm shift in precision medicine. The efficacy of this novel methodology in discerning molecular markers linked to distinct subtypes of cancer is supported by data demonstrating its high sensitivity, specificity, accuracy, precision, and F1 score. The comprehensive feature analysis provides insights into the molecular mechanisms underlying cancer and potential biomarkers by investigating the complexity of genomic material. By employing ethically sound practices and utilizing datasets like the Cancer Genome Atlas (TCGA), this approach improves cancer diagnosis accuracy and broadens the comprehension of cancer heterogeneity. As genomics and AI continue to be integrated, these results highlight the potential of AI-powered cancer diagnostics as a pivotal tool in the pursuit of more individualized healthcare interventions. The potential ramifications for individualized treatment strategies and enhanced patient results are extensive.

## References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA. Cancer J. Clin., vol. 68, no. 6, pp. 394-424, 2018.

[2] Y. B. Zhang et al., "Combined lifestyle factors incident cancer and cancer mortality: a systematic review and meta-analysis of prospective cohort studies", Br. J. Cancer, vol. 122, no. 7, pp. 1085-1093, 2020.

[3] C. Y. Lin et al., "Deep learning with evolutionary and genomic profiles for identifying cancer subtypes", J. Bioinform. Comput. Biol., vol. 17, no. 3, pp. 1-15, 2019.

[4] E. A. Collisson, P. Bailey, D. K. Chang and A. V. Biankin, "Molecular subtypes of pancreatic cancer", Nat. Rev. Gastroenterol. Hepatol., vol. 16, no. 4, pp. 207-220, 2019.

[5] I. Subramanian, S. Verma, S. Kumar, A. Jere and K. Anamika, "Multi-omics Data Integration Interpretation and Its Application", Bioinform. Biol. Insights, vol. 14, pp. 7-9, 2020.

[6] S. Graw et al., "Multi-omics data integration considerations and study design for biological systems and disease", Mol. Omi., vol. 17, no. 2, pp. 170-185, 2021.

[7] O. Menyhart and B. Gyorffy, "Multi-omics approaches in cancer research with applications in tumor subtyping prognosis and diagnosis", Comput. Struct. Biotechnol. J., vol. 19, pp. 949-960, 2021.

[8] C. H. Zheng, D. S. Huang, X. Z. Kong and X. M. Zhao, "Gene Expression Data Classification Using Consensus Independent Component Analysis", Genomics Proteomics Bioinforma., vol. 6, no. 2, pp. 74-82, 2008.

[9] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma and Sandeep Kaushik, "Big data in healthcare: management analysis and future prospects", Journal of Big Data, 2019.

[10] Usman Akhtar, Jong Won Lee, Hafiz Syed Muhammad Bilal, Taqdir Ali, Wajahat Ali Khan and Sungyoung Lee, "The Impact of Big Data In Healthcare Analytics", 4th International Conference on Information Networking (ICOIN), 2020.

[11] A. R. Bhat and R. Hashmy, "Artificial Intelligence-based Multiomics Integration Model for Cancer Subtyping," 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2022, pp. 536-539.

[12] H. T. Sadeeq, S. Y. Ameen and A. M. Abdulazeez, "Cancer Diagnosis based on Artificial Intelligence, Machine Learning, and Deep Learning," 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakheer, Bahrain, 2022, pp. 656-661.

[13] E. Weitschek, F. Cumbo, E. Cappelli and G. Felici, "Genomic Data Integration: A Case Study on Next Generation Sequencing of Cancer," 2016 27th International Workshop on Database and Expert Systems Applications (DEXA), Porto, Portugal, 2016, pp. 49-53.

[14] A. Singh and S. Kumar Jain, "A Personalized Cancer Diagnosis using Machine Learning Models Based on Big Data," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020, pp. 763-771.

[15] V. Iyer, A. M. Hima Vyshnavi, S. Iyer and P. K. K. Namboori, "An AI driven Genomic Profiling System and Secure Data Sharing using DLT for cancer patients," 2019 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2019, pp. 1-5.