Exploring DNA Methylation Biomarkers and Deep Learning for Cancer Epigenetics

Remyamol K M[1] and Philip Samuel[2]
[1]Department of Information Technology, School of Engineering, CUSAT, Kerala, INDIA
[2]Department of Computer Science, CUSAT, Kerala, INDIA

[1]Corresponding Author: rems84@gmail.com

Received: 01-05-2024 Revised: 17-05-2024 Accepted: 04-06-2024

ABSTRACT

Changes in DNA methylation, such as overall reduced methylation and increased methylation in specific CpG islands, are commonly seen in different types of cancer and also used as biomarkers to detect and diagnose cancer at an early stage. The distinct DNA methylation patterns provide valuable information about cancer progression and therapy. Recent advances in high-throughput techniques such as genome-wide profiling have transformed epigenetics by allowing computational analysis of intricate DNA methylation data. Deep learning techniques have become effective instruments for examining these patterns of methylation, enabling the identification of cancer markers, categorization of tumors, filling in missing data, and forecasting patient survival. This comprehensive review investigates the various uses of deep learning in examining DNA methylation and multi-omics data for cancer studies. It presents state-of-the-art deep learning architectures that are capable of addressing obstacles linked to research on cancer epigenetics. Nevertheless, the review also recognizes possible restrictions and areas for future investigation in this swiftly developing field. This work seeks to improve cancer diagnostics and therapeutic strategies by tackling these challenges and furthering knowledge of epigenetic mechanisms in cancer.

***Keywords*--** DNA Methylation, Cancer Epigenetics, Deep Learning, Biomarkers, Machine Learning, Cancer Diagnosis

I. INTRODUCTION

Recent improvements in computational techniques have been incorporated into medical research, leading to a rise in the generation of genomic data that offers both challenges and possibilities for disease detection and treatment. The development of high-throughput technologies has changed how genomic data is obtained, but the large amount and complexity of the data require advanced infrastructures and bioinformatics pipelines for successful management, storage, analysis, and security. Preprocessing is key in this situation, particularly feature selection, to improve data quality and ensure the precision and reliability of upcoming analyses. While cancer is often associated with genetic mutations, the significance of epigenetics, specifically DNA methylation, is increasingly acknowledged in the progression of illnesses. Epigenetic modifications, serving as a secondary layer of genetic data, intricately regulate gene expression and cell functionality, offering important insights into cancer biology. Using alterations in DNA methylation as biomarkers has significant potential for enhancing early detection of cancer, supporting treatment choices, and facilitating disease monitoring. However, there are still major challenges when analyzing genomic data, primarily because of its intricate nature that makes traditional statistical methods difficult to apply [1] [2]. Dimensionality reduction methods are necessary for solving this issue as they enhance interpretability and computational efficiency while retaining key information. Moreover, feature selection methods are essential in cancer genomics, assisting in the search for biomarkers and customizing treatment by pinpointing important genes and interpreting tumor characteristics within the data. The revolutionary shift in precision medicine is underscored by the intricate interplay between advanced computational techniques, genomic data analysis, and clinical insights [3].

The growth of computational techniques has led to an increase in genomic data generation, presenting challenges in disease detection and management. Complex data generated by high-throughput technologies requires sophisticated management and analysis for effective handling. Choosing features is crucial in order to ensure the dependability of data. Even though cancer has typically been associated with genetic changes, the significance of epigenetics, particularly DNA methylation, is increasingly recognized for its role in cancer progression. Alterations in DNA methylation serve as possible indicators for early detection and guidance of therapy [4][8]. However, the examination of genomic data is hindered by its high number of dimensions, necessitating the utilization of methods to decrease dimensionality. Choosing the right methods to select features is essential in pinpointing significant genes and tumor subtypes. The convergence of computational techniques, genomic examination, and clinical understanding signals the start of a fresh era in precision medicine.

DL, a specialized branch of ML, operates similarly but with enhanced capabilities, utilizing complex algorithms and network setups to extract insights from data. This systematic review focuses on the use of deep learning in cancer epigenetics, particularly with datasets like DNA methylation data. These research projects employed different deep learning structures to explore cancer biomarkers, categorize samples, forecast methylation statuses, and evaluate patient survival. The test provides details about novel advancements in DL-driven approaches for researching cancer epigenetics, addressing challenges and showcasing potential applications in medical settings.

II. MATERIALS AND METHODS

The process of DNA methylation, which adds methyl groups to cytosine bases, is essential for controlling gene expression and has important epigenetic implications, especially in conditions such as cancer. Multiple techniques such as bisulfite sequencing and immunofluorescence staining have been devised for the analysis and quantification patterns of DNA methylation. While initial methods looked at methylation levels in general, newer approaches allow for the analysis of methylation at particular sequences and areas of the genome. Despite the increasing curiosity in DNA methylation as a possible biomarker for early cancer detection and diagnosis, the transition to clinical settings is hindered by technical constraints and financial factors. This review emphasizes the significance of DNA methylation markers in the detection and prediction of cancer, exploring both current and developing markers, challenges in technology, and the opportunity for big data and innovative technologies to shape future methylation biomarkers [5] [6]. Moreover, the review delves into how artificial neural networks (ANNs) are used to analyze complex methylation and gene expression data in tumor profiles, helping to uncover cancer subtypes and patterns. In general, the review highlights the importance of analyzing DNA methylation and the promise of advanced computational methods in pushing forward cancer research and precision medicine. The work flow of the study is shown in Fig.1.

Convolutional neural networks (CNNs) have transformed computer vision by showcasing exceptional abilities in tasks like image classification, object detection, and segmentation. Their structured layout enables them to autonomously grasp and identify complex details in unprocessed information, making them very efficient at handling visual data. CNNs use convolutional layers to apply trainable filters to input images, detecting patterns and features of various sizes within the images. Pooling layers decrease the size of feature maps, diminishing computational load while maintaining important data. Fully connected layers delve deeper into the extracted features, allowing for tasks such as classification or regression to be performed. CNNs have been widely used in different fields apart from just computer vision, such as natural language processing, bioinformatics, and medical imaging, because of their capacity to understand intricate patterns from data with many dimensions [7].

Deep autoencoders (AEs) are used as models for unsupervised learning, with the goal of efficiently encoding and decoding high-dimensional data. AEs are useful for reducing dimensions, cleaning data, and extracting features by understanding compressed versions of input data that can recreate the original information with little loss. The encoder component of an autoencoder converts input information into a reduced-dimensional latent space representation, while the initial input was rebuilded by decoder using this representation [9]. AEs have been utilized in various industries, such as image compression, anomaly detection, and generative modeling, because they can extract important information from unprocessed data without needing supervision.

Capsule networks, also known as CapsNets, offer a new way to address the drawbacks of conventional convolutional structures, especially in terms of maintaining spatial details during max pooling. CapsNets use capsules to contain details about particular entities or components of objects in an image, aiding in hierarchical feature extraction and maintaining spatial connections. Within CapsNets, the interactive routing system enables capsules to communicate and agree on object presence and position, enhancing object recognition and reassembly capabilities. This structure provides exciting progress in areas like object recognition, pose estimation, and image segmentation, by obtaining more detailed and organized visual data representations.

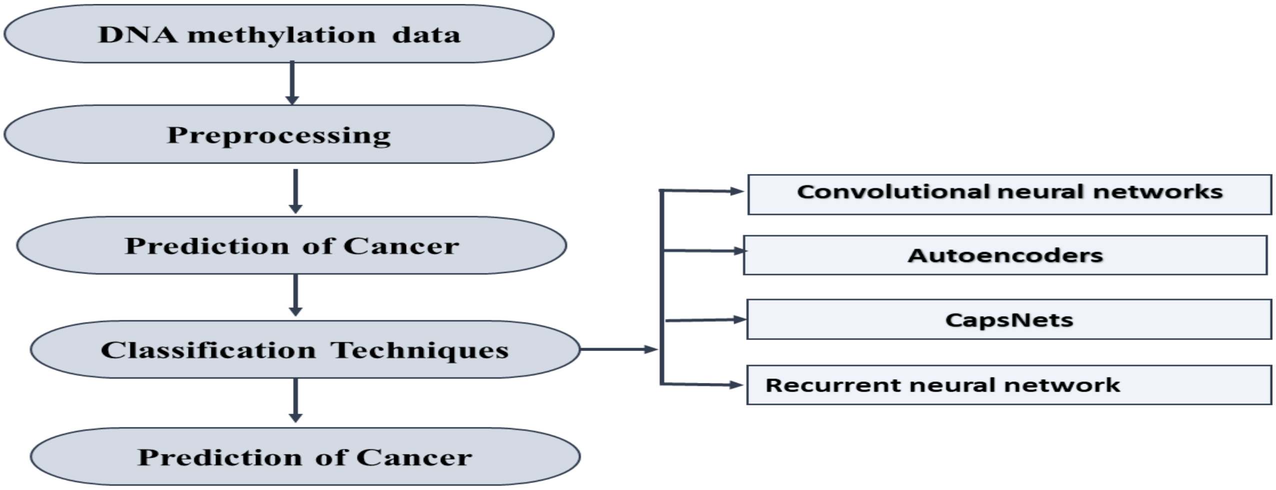

Figure 1: Work Flow diagram

RNNs are specifically created to analyze sequential data by including loops that enable information retention across time. This makes them perfect for tasks like time series forecasting, natural language processing, and speech recognition where the sequence of input elements is important for comprehension and prediction. LSTM units, a type of RNNs, solve the problem of disappearing gradients and help with capturing long-term relationships in sequential data [4] [11]. LSTMs have been used in language modeling, sentiment analysis, and music generation, due to their capability to capture temporal dynamics and memory. GRUs provide a simpler option to LSTMs with less parameters and gates, allowing for efficient computation while maintaining the capability to represent sequential relationships efficiently. GRUs have been utilized in tasks like machine translation, video analysis, and speech synthesis, where the need for real-time processing and memory efficiency is crucial.

Deep learning methods show great potential in understanding the intricate connections between epigenetic patterns and disease states, specifically in the field of cancer research, within the realm of DNA methylation analysis. Utilizing advanced deep learning techniques like CNNs, AEs, CapsNets, and RNNs allows researchers to efficiently analyze extensive methylation datasets, pinpoint biomarkers specific to certain diseases, and reveal fresh understandings of disease mechanisms. These advanced computational methods allow for the merging of various types of data, including gene expression, methylation, and clinical information, to improve disease diagnosis and prognosis with greater precision and thoroughness [12]. Even though noisy and high-dimensional methylation data present difficulties, deep learning algorithms provide strong solutions for extracting features, recognizing patterns, and modeling predictions, advancing personalized medicine and precision oncology efforts. While scientists keep improving deep learning techniques, there is a lot of potential for significant advancements in epigenetics and cancer research, offering new possibilities for early detection, precise treatment, and better results for patients.

III. RESULTS AND DISCUSSION

The study provided a complete exploration of device learning strategies implemented to most cancers diagnosis, survival evaluation, and treatment recommendation using DNA methylation records sourced from The Cancer Genome Atlas (TCGA). The study applied a numerous neural community models, accomplishing remarkable overall performance metrics, which includes a 100% area under the curve (AUC) and 94% accuracy, with precision and sensitivity both at 97%. To visualize disease clusters, the predictions of device mastering models were transformed right into a two-dimensional area by dimensionality reduction techniques. The dataset encompassed DNA methylation records from 485,000 sites, which includes 9,090 most cancers samples and 746 regular samples across 33 most cancers types, supplemented by scientific facts from TCGA. For most cancers diagnosis, numerous ML models consisting of Support Vector Machine, Multi-Layer Perceptron, Naive Bayes and k-Nearest Neighbor, were built, with the latter optimized the use of focal loss and cross-entropy loss functions. Survival analysis concerned the identity of 145 methylation markers predictive of survival outcomes, with the development of a prognostic model primarily based totally on a combined prognostic score (cp-score) to stratify sufferers into high- and low-threat groups. Treatment advice models, constructed the use of NB, KNN, SVM, and MLP, addressed class imbalance the use of the Synthetic Minority

Over-sampling Technique (SMOTE). Additionally, feature selection algorithms consisting of Wilcoxon rank-sum test, Benjamini-Hochberg procedure, LASSO, and Random Forest have been employed to identify considerable methylation sites and markers for diagnosis. The study`s predictive modeling employed logistic regression, SVMs, random forests, and naive Bayes classifiers, with the Matthews correlation coefficient (MCC) serving because the number one evaluation metric, complemented by permutation testing to evaluate statistical significance [8] [13] [14]. Overall, the study showcased the efficacy of machine learning approaches offering valuable insights into personalized oncology and precision medicine initiatives. The prediction quality for our trials is assessed using various metrics: median absolute error (MAE) as well as mean squared log Error (MSLE). The median absolute error is given as:

$$\mathrm{MAE} = \frac{1}{x}\sum_{i=1}^{n}[x_{i-}y_i]$$

The equation calculates accuracy by comparing predicted class (xi) to actual class (yi) for each sample (n) in test set. The mean squared log Error is given as:

$$\mathrm{MSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\ (x_{i-}y_i)^2}$$

Additionally, we evaluate the classification results using the classification accuracy as well as F-measure metrics. The percentage of correctly predicted classes to all analyzed specimens is known as accuracy, as defined as:

$$\mathrm{Accuracy} = \frac{t_p+t_n}{t_p+f_p+t_n+f_n}$$

Thus, false negative (f_n) and false positive (f_p) represent the misclassified instances while true positive (t_p) and true negative (t_n) values refer to correct classification of positive and negative instances, respectively. When there are about equal numbers of examples for all groups and the cost of misclassifying cancer was high, then the accuracy works well enough. Therefore, also use F-measure for depicting classifier’s accuracy along with its reliability**.** The F-measure is calculated as in Eq. (4). As the F measure rises, the model's accuracy gets better. A balancing F-measure seems to be a single score since accuracy and recall have quite a harmonic mean that may have either value from zero to one.

$$\mathrm{F} = \frac{\mathrm{2xPrecisionxRecall}}{\mathrm{Precision+Recall}} = \frac{2t_p}{2t_p+f_p+f_n}$$

The research delved into an intricate analysis of machine learning (ML) techniques applied to cancer diagnosis, survival prognosis, and treatment recommendation, specifically leveraging DNA methylation data retrieved from The Cancer Genome Atlas (TCGA). Through the implementation of a deep learning (DL) network featuring three consecutive layers, the study achieved remarkable performance metrics, including an accuracy of 96%, with both precision and sensitivity reaching 97%. To visualize disease clusters effectively, the study adopted the transformation of Deep Learning and Machine Learning model predictions into a two-dimensional space using dimensionality reduction. The dataset utilized comprised DNA methylation data from a vast array of 485,000 sites, spanning 33 cancer types and incorporating 9,090 cancer samples alongside 746 normal samples sourced from TCGA. Complementing the genomic data, the study augmented its analysis with comprehensive clinical information also obtained from TCGA, encompassing patient demographics, treatment histories, and outcomes. For the task of cancer diagnosis, the study employed a diverse set of ML models with optimization [13]. Survival analysis was a key focus, involving a rigorous process of identifying 145 methylation markers predictive of survival outcomes. This process entailed splitting the dataset into training and validation sets, utilizing proportional hazards models for feature selection, and constructing a prognostic model based on a combined prognostic score.

In parallel, the study explored ML models for treatment recommendation, leveraging clinical data encompassing treatment records and disease recurrence information for 1,377 patients [14]. Feature selection algorithms, including Wilcoxon rank-sum tests, the Benjamini-Hochberg procedure, LASSO, and Random Forest, were harnessed to identify significant methylation sites and markers conducive to accurate diagnosis. Among these methods, the Random Forest approach was particularly effective, facilitating the screening of relevant features and the selection of the most informative markers. The predictive modeling were deployed, with the Matthews correlation coefficient (MCC) serving as the principal evaluation metric. Additionally, the study incorporated permutation testing to ascertain the statistical significance of observed predictive performance, providing empirical validation of the robustness and reliability of the developed models. Overall, this comprehensive investigation exemplifies the power of ML methodologies in harnessing DNA methylation data for a spectrum of clinical applications, ranging from accurate cancer diagnosis to personalized treatment recommendations and prognosis assessments, thereby advancing the forefront of precision oncology and personalized medicine initiatives.

IV. CONCLUSION

Our study explored cancer epigenetics, specifically examining the use of DNA methylation data in cancer research. We emphasized the importance of changes in DNA methylation in the development of cancer, underscoring their potential as biomarkers for early detection and personalized treatment approaches. We explored the latest progress in deep learning (DL) methods and how they are used to analyze DNA methylation data. These techniques in deep learning show potential opportunities for detecting cancer, categorizing it, finding biomarkers, and forecasting patient outcomes and cancer types. Nevertheless, we also tackled the difficulties related to analyzing DNA methylation data using deep learning methods. Challenges like interpretability, overfitting, imbalanced data, and limitations in resource availability must be addressed for successful execution in clinical environments. Even with these obstacles, we highlighted the significance of being transparent in attaining clinical precision and described new DL tools that could transform upcoming DNA methylation studies. In conclusion, our study offers a thorough examination of the present status and future opportunities for employing DL techniques in the analysis of DNA methylation data in cancer research, with the goal of advancing early detection and personalized treatment methods.

REFERENCES

[1] Nakagawa, H. & Fujita, M, (2018). Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci., 109*(3), 513–522.

[2] Davalos V & Esteller M. (2023). Cancer epigenetics in clinical practice. *CA Cancer J Clin*. DOI: 10.3322/caac.21765.

[3] Merkel A & Esteller M.(2022). Experimental and bioinformatic approaches to studying DNA methylation in cancer. *Cancers (Basel)*. DOI: 10.3390/cancers14020349.

[4] Baldi P. (2012). Autoencoders, unsupervised learning, and deep architectures. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*.

[5] Williams RJ & Zipser D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computing*.

[6] Chatterjee A & Eccles MR. (2015). DNA methylation and epigenomics: new technologies and emerging concepts. *Genome Biology*. DOI: 10.1186/s13059-015-0674-5.

[7] Chatterjee A, Macaulay EC & Rodger EJ. (2016). Placental hypomethylation is more pronounced in genomic loci devoid of retroelements. *G3 Genes Genomes Genet*.

[8] Meissner A, Gnirke A & Bell GW. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.*

[9] Merkel A & Esteller M. (2022). Experimental and bioinformatic approaches to studying DNA methylation in cancer. *Cancers (Basel)*. DOI: 10.3390/cancers14020349.

[10] Asada K, Kaneko S & Takasawa K. (2021). Integrated analysis of whole genome and Epigenome data using machine learning technology: toward the establishment of precision oncology. *Front Oncol*.

[11] Krittanawong C, Zhang H & Wang Z. (2017). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology*.

[12] LeCun Y, Bengio Y & Hinton G. (2015). Deep learning. *Nature*. DOI: 10.1038/nature14539.

[13] Krizhevsky A, Sutskever I & Hinton GE. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*.

[14] Baldi P. (2012). Autoencoders, unsupervised learning, and deep architectures. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning, PMLR*.