Deciphering Cancer's Code: A Review on Advanced Machine Learning Approaches in DNA-Based Cancer Detection

Dr. Shyamal S. Virnodkar[1], Dr. Sangita B. Nemade[2], Varsha P. Gaikwad[3], Shreya Shinde[4], Purvi Shah[4], Naman Dagade[4]

[1]*Associate Professor, Department of Computer Engineering, K. J. Somaiya Institute of Technology, India.*
[2]*Assistant Professor, Computer Engineering Department, Government College of Engineering and Research, Avasari, Pune, India.*
[3]*Assistant Professor, Information Technology Department, Government College of Engineering, Aurangabad, Maharashtra, India.*
[4]*Bachelor of Technology, Department of Computer Engineering, K.J. Somaiya Institute of Technology, India.*
Email: shyamal@somaiya.edu,

KEYWORDS
Cancer,advanced machine learning.

ABSTRACT
This review provides a comprehensive synthesis of the latest developments in cancer detection and classification through the application of deep learning and machine learning techniques. An extensive range of methodologies is thoroughly assessed in the paper, encompassing Principal Component Analysis, Singular Value Decomposition, Autoencoders, Deep Belief Networks, Convolutional Neural Networks, and numerous architectures for cancer detection. The clinical implications and transformative potential of these computational approaches to improve the accuracy and efficacy of cancer diagnosis are highlighted. The paper presents an orderly supposition of the investigations conducted in the field, wherein each study introduced distinct approaches and methodologies aimed at augmenting early cancer detection and educating approaches to treatment. Although the results are encouraging, the article stresses the importance of conducting more extensive validations on a wide range of patient populations and investigating potential synergies with complementary technologies. The results obtained from these investigations represent substantial advancements in the field of biomedical informatics, offering innovative approaches that have the potential to revolutionize personalized medicine and healthcare.

1. Introduction

Cancer, a complex series of disorders distinguished by unusual cellular proliferation and propensity for invasion, presents a substantial healthcare dilemma. Malignant tumours, characterised by uncontrolled cellular proliferation and the capacity to metastasize, necessitate a comprehensive strategy for both identification and management. Early detection is critical for achieving effective intervention and an improved prognosis. In contrast, early warning signs, including chronic coughs, irregular blood loss, fluctuations in body weight, and changes in physiological patterns, frequently intersect with symptoms of alternative medical conditions, thereby impeding prompt diagnosis. The complexities of cancer, which encompass more than one hundred unique subtypes, arise from impairments in the integrity of DNA, specifically impacting genes that control cell division and growth. Environmental factors, such as specific viral agents, carcinogenic substances, and ionising radiation, have the potential to induce genetic mutations.

In light of these obstacles, novel technological resolutions have surfaced, with Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) emerging as particularly potent instruments in the realm of cancer management. These technological advancements examine complex datasets, including diagnostic imagery and electronic health records, uncovering nuanced patterns and correlations that are associated with cancer. AI algorithms have exhibited superior precision in detecting cancerous anomalies in medical images compared to human radiologists [1]. In addition, artificial intelligence models are making advancements in the direction of offering personalised treatment suggestions that are predicated on the distinct genetic composition of each individual. The application of AI, ML, and DL to lung cancer, the primary cause of cancer-related mortality, is illustrative. The utilisation of these technologies has greatly improved the precision of early detection through the analysis of complex patterns in medical visualisations, exceeding the limitations of human perception [2]. Additionally, through a careful examination of genetic characteristics and lifestyle decisions, AI-driven approaches have the capability to detect individuals who are susceptible to developing lung cancer.

Although AI, ML, and DL are still in their infancy as integrators in cancer diagnostics and prognostics, their potential for revolutionising the field is indisputable. This article examines the dynamic field of artificial

intelligence (AI) implementations in the identification of cancer, concentrating on pulmonary cancer. The objective is to emphasise the transformative capacity of these technologies in advancing patient outcomes and influencing the trajectory of cancer treatment. Moreover, in the realm of cancer research, AI's influence transcends clinical applications by expediting the discovery of innovative biomarkers, prospective pharmaceutical candidates, and therapeutic targets. By sifting through extensive datasets, AI-powered tools accelerate the drug discovery process, potentially reducing the time between the laboratory bench and the patient's bedside and providing hope to countless patients awaiting innovative therapies.

The objective of this review is to present a thorough and inclusive analysis of the present condition of the field, with particular attention given to the significant advancements achieved in the diagnosis and treatment of lung cancer. The significance of AI, ML, and DL in shaping the trajectory of cancer care is emphasised, as they not only provide accurate treatment and timely detection but also instill renewed optimism regarding improved results for patients.

DNA based Cancer Detection Methods

Using ML and DL to find cancer in DNA is a cutting-edge method that takes advantage of the huge amount of genetic data that is now available. ML and DL models can find genetic mutations or markers linked to different types of cancer by looking at DNA patterns. This makes cancer detection very accurate and early. These complex algorithms can find small patterns and oddities in genetic information. This makes personalised medicine and better cancer treatments possible.

I. Supervised Learning Methods

DNA-based cancer diagnosis uses supervised learning to look at genetic information and find biomarkers related to cancer. Labelled datasets are used in these methods to connect certain genetic patterns with the presence or forms of cancer. These methods help make accurate cancer diagnoses and classifications by training and validating machine learning models like support vector machines (SVM) or random forests (RF). This leads to earlier spotting and more personalised treatment plans.

1. Linear Models

Support Vector Machine(SVM):

The role of ML and SVMs has been nothing short of transformative in the field of cancer study, which is always changing. These strong formulas have changed the field over the years, making it easier to find and diagnose cancer. Looking at important turning points and successes in cancer research, along with the newest and most cutting edge ideas of 2023, as SVMs and other new methods continue to make their mark in the fight against cancer.

At the start of the 2000s, Brown et al. [3] made a big discovery in the field of cancer study. The writers used SVM to sort microarray data and were able to tell the difference between cancer cells and normal cells with over 90% accuracy. The author's SVM model did better than other methods, showing that the SVM is good at figuring out complicated patterns in high-dimensional microarray datasets. Guyon et al. [4] changed the field when they created Recursive Feature Elimination (RFE) for cancer classification with SVMs in 2002. This method for ranking genes found a small group of genes, which made the SVM predictor much more accurate and showed that SVMs could be used in cancer research.

In 2003, Valentini et al. [5] showed off their clever way to find cancer. The author's method used bagged groups of SVMs, which made it a lot more accurate. DNA microarray data from breast cancer patients and healthy controls were used to test it. It did very well (93.3%), beating individual SVMs.

In 2015, Guo et al. [6] used an SVM algorithm to find a group of DNA methylation biomarkers for NSCLC. This showed hope for creating a very accurate diagnostic test for this deadly disease. In 2022, Ray Bahado-Singh et al. [7] showed a potential way to find lung cancer early by combining an SVM classifier with AI and DNA methylation analysis. The author's SVM helped tell the difference between lung cancer cases and healthy controls. This shows that SVM-based ML combined with DNA methylation analysis could help doctors find lung cancer earlier. In their 2023 study, Hosseini et al. [8] looked at DNA methylation and gene expression data from people with stomach cancer as a whole. The writers found 256 DMPs and 2821 DEGs. The writers made a very good diagnostic model using the SVM classifier. In addition, the study found four DNA methylation probes that are linked to total survival. These results show good biomarkers for diagnosing and predicting

stomach cancer, which could have effects on finding the disease early and telling how well a patient will do. In 2023, Venkatesan N et al. [9] came up with a useful way to find lung cancer using SVM that was improved by an Improved Weight-based Beetle Swarm Optimisation (IW-BS) method. The results are about the same as those of modern methods, which shows that SVM with IW-BS could be a good way to find lung cancer early.

Logistic Regression(LR):

The importance of early cancer detection has been heightened in light of recent advancements in medical and genetic technologies, with the intention of enhancing patient outcomes and decreasing cancer-related fatalities. Chen et al. [10] pioneered the development of a methylation-based circulating tumour DNA (ctDNA) assay, which presents a viable alternative to invasive methods in the identification of various forms of cancer. Particularly with regard to cases in their advanced stages, the research revealed a remarkable degree of precision in the detection of cancer. Furthermore, Ahlquist et al. [11] unveiled an innovative stool DNA test of the subsequent iteration designed to screen for colorectal cancer (CRC). Their study underscored the potential of stool DNA as a non-invasive instrument for detecting CRC by analysing methylated DNA and mutant KRAS DNA.

In 2003, Sozzi et al. [12] established an innovative contribution when they demonstrated that free circulating DNA (cfDNA) could function as a diagnostic marker for the early detection of lung cancer, providing an exceptionally precise and non-invasive method. In their investigation of urine DNA methylation as an early diagnostic marker for bladder cancer, Hoque et al. [13] achieved an overall accuracy of 89% in detecting the disease.Jensen et al. [14] investigated hypermethylated DNA as a prospective biomarker for non-invasive screening in the context of colorectal cancer. Bach et al. [15] further elaborated on this by monitoring urine samples for DNA methylation and detecting two markers, SEPT9 and SDC2. In an effort to improve the early detection of ovarian cancer, Wu et al. (2019) [16] utilised DNA methylation biomarkers extracted from cervical scrapings. Imperiale et al. [17] published a description of multitarget stool DNA testing that holds promise for the detection of colorectal cancer (CRC) at an early stage.

Chalasani et al. (2021) [18] distinguished themselves by introducing a panel of methylated DNA and protein markers for early-stage detection of hepatocellular carcinoma (HCC) in the blood. These investigations utilise LR to illuminate the potential of blood-based markers, DNA methylation, cfDNA, and stool DNA in transforming non-invasive cancer detection. These innovative techniques hold great potential for enhancing patient care and reducing cancer-related fatalities.

2. Tree based Models

Decision Tree:

In their respective investigations, Chen et al. [19], José M. Jerez-Aragonés et al. [20], and J. Keerthika et al. [21] explored innovative methodologies in the field of breast cancer research. Chen et al. [19] utilised a decision tree algorithm to examine potential associations between DNA viruses and breast cancer. Particular pathogens associated with the illness were identified by the author's model, emphasising potential repercussions. A novel integrated neural network and decision tree model was devised by José M. Jerez-Aragonés et al. [20] with the purpose of forecasting breast cancer relapse. The sensitivity and specificity of the author's model were also remarkable, achieving an extraordinary 76.2% accuracy. J. Keerthika et al. [21] devised a data mining methodology that identified breast cancer with an extraordinary 98.25% precision by utilising decision trees.

In aggregate, these investigations highlight auspicious avenues for the investigation, diagnosis, and prognosis of breast cancer, advocating for more comprehensive assessments of validation and clinical applicability.

Random Forest (RF):

In the domain of lung cancer diagnosis and research, RF, an ensemble learning algorithm, has emerged as a potent instrument in the last decade. A multitude of scholarly investigations have examined its capacity to augment the early identification, categorization, and prognostication of lung cancer instances. This narrative elucidates the substantial contributions made by various research teams in the domain of RF-based healthcare applications by interweaving the results of these studies. Ahuja et al. [22] introduced the MethyLight technique in 2008, allowing for the investigation of DNA methylation patterns in non-tumorous lung tissues and squamous cell lung cancer (SCLC). By identifying a panel of eight hypermethylated markers, the authors paved the way for the detection of cancer via DNA methylation, specifically SCLC. Expanding upon this groundwork,

Ponomaryova et al. [23] investigated the potential of aberrantly methylated ctDNA as a non-invasive tool for lung cancer surveillance and diagnosis in 2013. As an adjunct to the capabilities of ML algorithms such as RF, the author's research revealed that lung cancer patients had substantially elevated methylation levels, which bodes well for early detection and monitoring. Further, in 2015, Cai et al. [24] combined SVM for classification with RF for feature selection from DNA methylation data. With its remarkable accuracy, the author's ensemble-based method offers optimism for the advancement of lung cancer diagnosis and treatment. In their 2017 article, Zou et al. [25] proposed a novel technique for determining the location of the primary tumour in cases of Cancer of Unknown Primary (CUP), which extends beyond lung cancer. By utilising RF in conjunction with tissue-specific markers, the authors predicted the origin of tumours with remarkable precision on a diverse dataset, thereby demonstrating the robustness and interpretability of RF. Ooki et al. [26] identified highly predictive biomarkers for lung adenocarcinoma (LUAD) utilising an RF classifier and DNA methylation profiling in 2019. The efficacy of the algorithm in identifying reliable biomarkers for cancer detection and prognosis was highlighted in the author's study, which offered encouraging prospects for patients with LUAD. Chen et al. [27] identified DNA methylation markers that were linked to occult lymph node metastasis (LNM) in non-small cell lung cancer (NSCLC), representing a substantial advancement in the field. The remarkable accuracy attained by their RF model has the potential to revolutionise approaches to early detection and treatment of NSCLC involving occult LNM. In the year 2020, Wang et al. [28] introduced an innovative approach by integrating self-paced bootstrap sampling and RF ensemble learning to enhance the accuracy of prognostications based on gene expression data for lung cancer. The success of their method over pre-existing approaches underscores its viability for additional verification using more extensive datasets. Pratama et al. [29] introduced a novel method for classifying microarray data by combining clustering and RF in 2018. Standard RF was surpassed in performance by the author's method, which holds the potential to enhance the classification of microarray data through the identification of informative features while reducing redundancy. Annisa et al. [30] contrasted feature selection methods and an RF classifier for cancer detection using microarray data in one of the most recent studies published in 2023. The results of their study highlighted the potential of LASSO as an effective method for selecting features, specifically in relation to lung cancer. Wang et al. [31] continued the exploration of novel biomarkers for early lung cancer detection via LC-MS/MS and plasma samples in 2023. This was the third time they had utilised ML methods. In order to identify differentially expressed metabolites, the RF algorithm has proven effective; however, additional validation in larger cohorts is required. RF, an ensemble learning algorithm, has been instrumental in lung cancer research for the past decade. Its capabilities have been harnessed in numerous studies to aid in prognosis prediction, classification, and early detection, with a particular emphasis on DNA methylation analysis, feature selection, and microarray data classification. Utilising RF's capabilities, scientists have identified biomarkers, predicted the origins of tumours, and enhanced the precision of cancer detection and prognosis. This adaptable algorithm remains a symbol of optimism in the continuous effort to eradicate lung cancer.

3. Instance Based Algorithm

K Nearest Neighbours (KNN):

KNN classification for DNA-based cancer detection is a valuable instrument in the field of oncology. By analysing genomic data, KNN can precisely predict medical outcomes and classify tumours into specific types of cancer. By capitalising on the potential for earlier detection and personalised treatment plans, this approach exploits the capabilities of machine learning to enhance cancer diagnostics, thereby significantly influencing the management of cancer.

Within the realm of cancer research, seminal contributions have surfaced, each introducing novel approaches to the diagnosis of the disease. Yang et al. [32] unveiled a non-invasive methodology for the detection of colon cancer in 2016. By employing Raman spectroscopy, Principal Component Analysis (PCA), and KNN, the researchers successfully categorised Raman spectra originating from healthy controls and patients diagnosed with colon cancer. The promising performance of the KNN classifier in terms of accuracy provides encouragement for the development of non-invasive, efficient early detection methods for colon cancer. In 2017, Li et al. [33] introduced a non-invasive FTIR spectroscopy-based method that advanced the field of CRC and colitis diagnosis. Enhanced KNN classifiers, which incorporated a feature selection algorithm and weighted distance metric, outperformed conventional KNN classifiers. Enhanced early diagnosis of colorectal cancer (CRC) and colitis may result from this innovation, which has the potential to improve patient survival rates.

In their study on leukaemia, Kumar et al. (2016) [34] utilised microarray data in conjunction with an effective KNN classifier based on MapReduce. This innovative method classified and identified subtypes of leukaemia patients with high precision, demonstrating the diagnostic potential of MapReduce and microarray data in the context of leukaemia. Bhuvaneshwari and Dr. Therese [35] made strides in the detection of lung cancer in 2023 through the utilisation of a genetic algorithm (GA) to optimise KNN classification. The approach they utilised, which was constructed using a dataset comprising data from both lung cancer patients and healthy controls, yielded exceptionally precise outcomes. This demonstrates the potential of KNN in conjunction with GA to enhance the diagnosis of lung cancer.

4. Neural Networks

Within the domain of advanced cancer research, a number of studies have introduced novel methodologies that hold the potential to profoundly influence the process of cancer diagnosis and classification.

In their study, Zheng and Xu (2022) [36] demonstrate the remarkable 99.72% accuracy with which deep neural networks (DNNs) can predict the origins of cancer. Using a vast dataset of DNA methylation data from over 7,000 patients representing 18 distinct cancer types, they trained their DNN model, which exhibited superior performance compared to traditional gene expression profiling techniques. In their article titled "Classification of DNA Microarrays Using Artificial Neural Networks and ABC Algorithm," González-Rivera et al. [37] present an innovative methodology. DNA microarray classification is accomplished through the combination of artificial neural networks (ANNs) and the artificial bee colony (ABC) algorithm. The utilisation of the ABC algorithm to identify informative genes enables the authors to attain exceptional classification accuracy when applied to a wide range of DNA microarray datasets. This development holds great potential for substantially propelling the field of biological research forward.

Ringnér and Peterson [38] undertook a comprehensive examination of ANNs in the context of cancer diagnosis in 2003, highlighting their consistently superior accuracy in classifying a wide range of cancer datasets. This demonstrates the capacity of ANNs to discern intricate connections within microarray data, thereby enhancing the diagnostic process for cancer. In addition, Gokberk et al. [39] present an innovative methodology for the classification of prostate tissue. By implementing an ANN that has been trained using prostate-specific antigen (PSA) levels and mitochondrial DNA (mtDNA) data, the authors attain a noteworthy accuracy rate of 92.9%. This methodology holds the potential to efficiently classify prostate tissue by utilising a small dataset.

In conclusion, Marchevsky et al. [40] introduce an innovative methodology for the categorization of specific lung cancer cell lines according to DNA methylation markers. The authors employ ANNs and linear discriminant analysis (LDA) to identify twenty DNA markers that are differentially methylated for the purpose of classifying NSCLC and small cell lung cancer (SCLC) cell lines. The ANN model demonstrates remarkable performance, attaining an accuracy rate of 95% and highlighting its capability to classify lung cancer cell lines reliably using DNA methylation markers. M. Vimaladevi and V. Karthikeyan [41] present a robust hybrid methodology that integrates ANNs and GAs for the purpose of augmenting cancer classification. Through the implementation of PCA to effectively reduce dimensionality and GAs to select pertinent features, the authors successfully attain a remarkable classification accuracy of 98.25% when applied to breast cancer data. The potential clinical applications of this novel hybrid approach are substantial, thereby enhancing the body of knowledge in the field of oncology.

Collectively, these investigations showcase the potential of cutting-edge neural networks and algorithms to enhance the accuracy and classification of cancer diagnoses, thereby propelling noteworthy progress within the domain of oncology.

5. Bayesian Models

Naive Bayes:

In a 2022 study, Ranjbar et al. [42] introduced an innovative approach to feature selection with the aim of enhancing cancer diagnosis through the utilisation of DNA microarray data. Combining GAs with the Naive Bayes classifier, their approach identifies genes that are critical for precise diagnosis. In order to select a subset, GAs employ Naive Bayes to evaluate the significance of the genes. The aforementioned technique exhibits superior diagnostic sensitivity and accuracy when applied to DNA microarray data concerning breast cancer compared to alternative methods. This underscores the promising prospects of GAs and ML in the domain of cancer diagnosis. Muhammad Shahbaz et al. [43] introduced a data mining-based cancer diagnosis system in

2012 that made use of Naive Bayes classification. Using gene expression data from patients with and without cancer to train the system, it distinguishes cancer patients with an impressive 95% accuracy. A.R. Dhobe et al. [44] assess three ML algorithms for disease prediction in a 2019 paper. At 98.5%, the Naive Bayes algorithm exhibits superior performance, followed by the KNN algorithm and the J48 algorithm. This highlights the potential of Naive Bayes classifiers in accurately predicting malignant diseases.

Bayesian Network:

Gevaert et al. [45] proposed a methodology in 2006 that utilised Bayesian networks to integrate clinical and microarray data in order to forecast the prognosis of breast cancer. The research, carried out using a dataset comprising 295 patients diagnosed with breast cancer, obtained an exceptional accuracy rate of 78%. This result surpassed the performance of models that exclusively relied on clinical or microarray data.

Roupret et al. [46] investigated urine analysis-based methodologies for predicting the recurrence of urothelial cell carcinoma (UCC) in 2008. When comparing methylation and microsatellite analyses, it was found that microsatellite analysis had a higher area under the curve (AUC) of 0.819 than methylation's 0.448. The authors identified a set of six indicators with an 86% accuracy rate in predicting UCC recurrence by utilising Bayesian network analysis. This provides valuable insights into the development of effective tools for predicting UCC recurrence. An innovative Dynamic Bayesian Network (DBN) model for cervical cancer screening was introduced by Oni'sko et al. [47] in 2008. This model was trained using a dataset comprising more than 100,000 women, both with and without cervical cancer. The predictive accuracy of this model for cervical cancer was found to be exceptional, surpassing that of conventional screening methods.

Roupret et al. [48] proposed a dynamic Bayesian network (DBN) framework in 2019 to classify time series microarray data for cancer. The DBN model demonstrated exceptional accuracy in the classification of cancer samples, outperforming traditional ML methodologies. Additionally, it unveiled gene associations with specific forms of cancer. This study underscores the potential of DBNs as robust instruments for precise and efficient gene discovery and cancer classification, with further investigation centering on validation and integration with complementary technologies.

6. Other Supervised Methods

Hidden Markov Model (HMM):

In their publication, Xiaoqing Yu and Shuying Sun (2015) [49] presented "HMM-DM," an innovative methodology that utilises Hidden Markov Models to detect differentially methylated regions (DMRs) within DNA methylation data. This epigenetic modification is crucial in the development of cancer and other diseases. Trained on bisulfite sequencing data from breast cancer patients, HMM-DM outperformed BSmooth, producing DMRs with greater accuracy and specificity. Significantly, it effectively identified DMRs that are linked to the advancement of breast cancer. Dehghani et al. [50] developed an HMM-based model for predicting the recurrence of breast cancer; it achieved an accuracy of 85%, outperforming conventional ML methods. In addition, this model identified loci associated with recurrence risk. These studies demonstrate the potential of HMMs to inform cancer treatment decisions and enhance our understanding of epigenetics.

Ensemble Learning:

Boosting Techniques:

AdaBoost: AdaBoost, an acronym for Adaptive Boosting, is a highly effective ensemble machine learning algorithm that has rendered substantial contributions to diverse domains of cancer diagnostics and research. The adaptability of this method and its ability to enhance the performance of base classifiers have attracted considerable interest from academia. Regarding tasks associated with cancer, AdaBoost has proven to be effective in numerous studies.

n recent times, machine learning algorithms, with a specific focus on Adaboost, have attracted considerable interest and exhibited exceptional potential across multiple domains of cancer diagnostics and research. Friedrich et al. (2012) [51] identified prostate cancer cell nuclei with Adaboost with an impressive 90% accuracy. This precision is comparable to that of other ML algorithms assessed in the research. The findings highlight the potential of Adaboost as a valuable instrument for accurately identifying the nuclei of prostate cancer cells, which is an essential component of DNA-grading in the evaluation of malignancy. Bhola and Tiwari [52] (2015) investigated the classification of cancers through the utilisation of gene expression data. Adaboost demonstrated

outstanding performance by accurately classifying gene expression data from breast cancer patients and healthy controls with an impressive 95% precision. The results of this study indicate that Adaboost exhibits considerable potential in the realm of cancer classification, surpassing alternative machine learning approaches in this particular domain.

Hajianfar et al. [53] (2019) utilised Adaboost to predict the status of O6-methylguanine-DNA methyltransferase (MGMT) in patients with glioblastoma multiforme (GBM) cancer using MRI radiomics features in a noninvasive manner. Adaboost accurately predicted the status of MGMT with a degree of 85 percent, a performance level comparable to that of other ML algorithms evaluated. The findings underscore the potential of Adaboost as a beneficial instrument for predicting the MGMT status noninvasively in patients with GBM, thereby presenting a hopeful pathway towards tailored treatment approaches. Nuhić et al. [54] (2020) investigated the potential of Adaboost and alternative classifiers in the context of ovarian cancer detection. Adaboost demonstrated a noteworthy 90% accuracy rate when it came to the classification of tumours using features extracted from ultrasound images. Consistent with the efficacy of other machine learning algorithms assessed in the research, this result validates the potential of Adaboost in the realm of ovarian cancer detection.

Uthamacumaran et al. [55] (2022) provided additional evidence of Adaboost's efficacy through the utilisation of vibrational spectroscopies to analyse extracellular vehicles (EVs) derived from cancer patients. Adaboost consistently exhibited a noteworthy 95% accuracy rate when classifying these electric vehicles, thereby outperforming alternative machine learning algorithms. The result highlights the potential of Adaboost as a valuable instrument for accurately characterising extracellular vesicles derived from cancer patients, an area that is gaining prominence in the field of cancer research.

In summary, Adaboost has demonstrated its versatility and potential as a machine learning algorithm across a range of cancer-related assignments, including but not limited to identifying cell nuclei, classifying gene expression, characterising extracellular vesicles, detecting tumours, and predicting MGMT status. The robust performance of this technology in a wide range of applications highlights its importance in furthering cancer research and diagnostics, which ultimately leads to enhanced clinical outcomes.

XGBoost: XGBoost, also known as Extreme Gradient Boosting, is an important algorithm in the field of ensemble machine learning (ML) and has demonstrated its exceptional utility in numerous domains of cancer diagnostics and research. This preamble establishes the context for the following segment concerning Adaboost through emphasising the importance of XGBoost.

Recent years have witnessed a profound revolution in cancer diagnostics and research, which has been propelled by the extraordinary capabilities of the XGBoost ensemble ML algorithm. The narrative unfolds via a sequence of investigations that collectively demonstrate the critical contribution of XGBoost to the progression of knowledge regarding cancer detection and classification.

It commences with [56] (2023), who proposed an innovative approach to forecast lung cancer through the examination of copy number variations (CNV) in plasma. By utilising XGBoost, the authors successfully attained an exceptional accuracy rate of 98%, thereby establishing a firm basis for a potentially effective tool in the early detection of lung cancer. Following this, Atlam et al. [57] (2020) proposed a feature selection method that utilises DNA microarray data for the purpose of cancer diagnosis. The authors identified crucial features utilising XGBoost and achieved a remarkable 95% accuracy in cancer diagnosis; this provides a novel avenue for the pursuit of precise disease identification. Swathi and Kodukula [58] (2022) contributed to the discourse by designing cancer predictions using XGBoost with Hyperband optimisation and gene expression data. The algorithm's remarkable accuracy of 96% underscored its potential as a dependable instrument for cancer prognosis. Ma et al. [59] (2020) contributed a chapter to this article by classifying cancers using multi-omics data and XGBoost with layering. The model proposed by the author, which achieved an impressive accuracy rate of 97%, underscored the significance of utilising a variety of data sources when classifying cancer.

Liu et al. [60] (2021) conducted an exhaustive assessment by integrating various omics data in order to deduce the tumour TOI. With a 97% accuracy rate when employing a multi-omics strategy, XGBoost demonstrated its efficacy in this critical area of cancer research. In their final work, Banjar et al. [61] (2022) present iDBP-PBMD, an innovative machine learning model designed to detect DNA-binding proteins. By utilising XGBoost and evolutionary profiles, the authors successfully attained a 96% accuracy rate, thereby introducing a potent instrument for the identification of critical DNA-binding proteins.

This section as a whole underscores the extraordinary trajectory of XGBoost as it revolutionises the domain of cancer diagnostics and research. XGBoost has demonstrated its versatility and indispensability as a tool in the continuous fight against cancer, facilitating insight into the tissue of origin, early detection, and accurate diagnosis. As a result, it offers improved patient outcomes and a more comprehensive comprehension of this intricate disease.

II. Unsupervised Learning Methods

1. Clustering:

A narrative that follows a chronological structure develops within the domain of cancer research, emphasising significant progressions and novel approaches towards enhanced diagnosis and treatment.

In 2001, Ramaswamy et al. [62] initiated an innovative endeavour by developing a method for multiclass cancer diagnosis through the utilisation of gene expression signatures. Their methodology successfully categorised various forms of cancer with an astounding 95% accuracy rate, thereby paving the way for data-driven cancer diagnostics. Simultaneously, Amor et al. [63] proposed a deep embedded refined clustering method that utilised DNA methylation data to accurately differentiate breast cancer subtypes. Their procedure guarantees patients with breast cancer individualised treatment plans.Concurrently in 2021, Lin Li and Xiaosheng Wang [64] introduced a novel methodology utilising pathway clustering to identify subtypes of gastric cancer. The three distinct gastric cancer subtypes identified by this novel method are distinguished by their own activation patterns of pathways, providing a promising avenue for the development of targeted therapeutic approaches.

In 2002, Virmani et al. [65] revolutionised the field by introducing hierarchical clustering as a method for classifying lung cancer cell lines based on DNA methylation profiles. The author's investigation unveiled four discrete clusters that corresponded to different subtypes of lung cancer, thereby offering a classification system and further understanding of the molecular distinctions among these subtypes. Gollapalle et al. [66] illuminated a critical facet of cancer progression in 2007 through the identification of oxidative clustered DNA lesions (OCDLs) in irradiated cells and tissues. An association between OCDLs and cancer was suggested by the author's method, which paved the way for the development of diagnostic biomarkers for early cancer detection. In the year 2018, Manogaran et al. [67] introduced a framework for large data processing based on machine learning that aimed to improve cancer prognosis and diagnosis. By employing a hidden Markov model and Gaussian mixture model clustering, they demonstrated exceptional precision in the identification of cancer subtypes and the prediction of patient prognosis. This has the potential to fundamentally transform the way cancer is managed.

In 2023, Sehhati et al. [68] contributed to the field by devising a technique for predicting the recurrence of breast cancer using gene expression data. The accuracy of the recurrence prediction achieved by the author through the integration of classification and k-means clustering techniques has the potential to significantly transform approaches to early detection and treatment of breast cancer recurrence. The systematic progression of these segments in cancer research establishes a cohesive storyline that encompasses precise diagnosis, identification of subtypes, early detection, and individualised treatment strategies. The potential for improved clinical outcomes and streamlined cancer management is encapsulated within this shared narrative.

2. Dimensionality Reduction

Principal Component Analysis (PCA):

In the attempt of developing more efficient methods for cancer detection and diagnosis, a number of novel strategies have surfaced. A method was proposed by Adiwijaya et al. [69] in which PCA is utilised to reduce the dimensionality of microarray data and uncover cancer-associated features. This method demonstrated a remarkable accuracy of 96.07% when evaluated on colon cancer data, surpassing conventional approaches such as support vector machines (SVMs) and logistic regression.

An additional significant development was made by D. Menaga and S. Revathi [70], who proposed a hybrid methodology that combined DL via neural networks with Probabilistic Principal Component Analysis (PPCA) for dimensionality reduction. The approach they employed exhibited exceptional potential in the domain of cancer categorization. By preprocessing gene expression data, reducing dimensionality with PPCA, and employing a deep neural network for classification, the authors utilised breast cancer gene expression data to achieve an astounding classification accuracy. In their investigation, Agarwal et al. [71] examined a novel

methodology that employed self-internalized plasmonic sensors to diagnose malignancy. When incubated with cells, these sensors, which are constituted of gold nanoparticles coated with DNA, interact with particular biomarkers associated with cancer. Traceable modifications in their plasmonic characteristics can be identified via the utilisation of a light microscope. Informed by PCA for dimensionality reduction, the author's method for detecting cancer cells attained remarkable precision.

An alternative method for screening for cancer was utilised by Harris et al. [72], who utilised Raman spectroscopy to analyse peripheral blood samples. The objective of the authors was to identify cancer-related alterations in blood samples using Raman spectroscopy. Sensitivity and specificity of their ML algorithm informed by PCA were exceptionally high when it came to classifying cancer patients. The utilisation of Surface-Enhanced Raman Spectroscopy (SERS) to detect esophageal cancer in tissue samples was investigated by Feng et al. [73]. By utilising PCA to reduce the dimensionality of the data and provide guidance to ML algorithms, their approach demonstrated remarkable sensitivity and specificity of 90.9% and 97.8%, respectively.

While these novel approaches provide optimism for enhanced cancer detection and diagnosis, additional research into complementary technologies and broader validation are necessary for optimising their capabilities.

Singular Value Decomposition (SVD):

Within the domain of biomedical research, novel approaches have been devised to tackle crucial challenges associated with the diagnosis and categorization of diseases. The novel paradigm for DNA microarray classification proposed by Huynh et al. [74] combines a feedforward neural network with SVD for dimensionality reduction in a two-step procedure. The method effectively addresses the challenges posed by high-dimensional DNA microarray data, achieving an exceptional classification accuracy of 95.3% in a dataset comprising breast cancer patients. This achievement surpasses that of current techniques. In the same way, Doufekas et al. [75] discuss the critical requirement for non-invasive and early cancer detection. The research investigated by the author explores the potential of DNA methylation as a marker specific to cancer. By analysing vaginal fluid samples from healthy individuals and cancer patients, the authors identify cancer-associated DNA methylation markers using SVD. The remarkable accuracy with which this panel of markers detects endometrial and cervical cancer highlights the promise of non-invasive screening tests based on DNA methylation.

Moreover, the research conducted by Jiang et al. [76] represents a significant progression in the utilisation of high-throughput DNA sequencing for CNV detection. By utilising SVD, their computational instrument, CODEX2, improves the precision of CNV detection. As a result of overcoming the inherent data biases and artefacts of high-throughput sequencing data, CODEX2 demonstrates its potential in cancer diagnosis, genetic counselling, and disease research with its exceptional sensitivity and specificity. In their final contribution to the field of breast cancer prediction, Junaid et al. [77] present an innovative method that is founded on DNA methylation analysis. By integrating SVD with vertical and horizontal DNA differential methylation analysis, they have developed a panel of predictive markers for breast cancer that exhibit an impressive sensitivity of 95% and specificity of 98%. The significance of this study lies in the potential of predictive tools based on DNA methylation to improve the early detection and diagnosis of breast cancer.

The aforementioned studies represent substantial advancements in the domain of biomedical informatics. They present encouraging approaches for the identification and categorization of diseases, while highlighting the need for additional verification and implementation in clinical settings.

3. Neural Network Based

Autoencoders:

The field of oncology research has undergone a substantial shift since the introduction of advanced ML methods, particularly the implementation of DL networks and autoencoders, which seek to improve the accuracy of cancer diagnosis and subtype categorization.

Hua et al. (2020) [78] conducted an exhaustive investigation into the capacity of various autoencoder architectures to identify cancer subtypes through the utilisation of multi-omics data. The results of their research emphasise the potential of autoencoders as highly effective tools for classifying complex cancer subtypes. In the field of oncology, Khwaja et al. (2019) [79] introduced an innovative method for subtype classification by utilising deep autoencoders to distinguish between various cancer types via unique DNA methylation patterns. This development significantly broadened the domain of subtype classification. Consistent with this, Xi Chen et

al. (2018) [80] demonstrated the transformative potential of DL models in the field of diagnostic applications by utilising deep autoencoders and neural networks to accomplish exceptional accuracy in cancer detection and type classification.

Zhang et al. (81) introduced a novel deep learning model called OmiVAE, which demonstrated its ability to integrate multi-omics data in a seamless manner to classify pan-cancer. The model in question possesses the capacity to fundamentally transform our approach to cancer classification. In addition, Adem (82) presented a novel methodology for the detection of breast cancer by employing a Stacked Autoencoder (SAE) in conjunction with Subspace kNN. This research suggests promising developments in the field of diagnostic techniques, ushering in an unprecedented era of breast cancer prognosis.

As a whole, these investigations underscore the tremendous potential that sophisticated machine learning methods possess for the identification and categorization of cancer, thereby inaugurating an age of individualised medicine. However, it is imperative to carry out additional comprehensive validations involving a wide range of patient populations and investigate possible synergies with supplementary technologies in order to maximise the potential of cancer diagnosis and classification systems.

Deep Belief Network (DBN):

Nayak et al. [83] utilised a DBN combined with Mutual Information (MI) in their endeavour to develop novel cancer detection methodologies. The authors evaluated microarray gene expression data with the primary objective of detecting breast cancer. Notably, their methodology achieved an exceptional accuracy rate of 94.74%, outperforming well-established techniques. In the interim, Ronoud and Asadi [84] investigated the diagnosis of breast cancer using an Extreme Learning Machine (ELM) in conjunction with an Evolutionary Deep Belief Network (E(T)-DBN). Across two datasets, this combination demonstrated exceptional accuracy of 96.87% and 94.36%, distinctly surpassing the performance of current methodologies. These innovative methodologies indicate the potential of DL in the field of cancer diagnostics; however, additional validation with larger and more diverse datasets is required.

III. Deep Learning Methods

Utilising DL methods for cancer detection from DNA has emerged as a cutting-edge strategy in contemporary oncology. In the analysis of intricate DNA data, DL algorithms, specifically Convolutional Neural Networks (CNNs), are utilised to detect mutations and patterns that are linked to various forms of cancer. By capitalising on the capabilities of DL, scientists and medical practitioners can attain enhanced precision and vividness in the identification of cancer stemming from DNA, thereby facilitating preemptive diagnoses and more efficacious therapeutic approaches. This novel application exhibits significant potential in the continuous effort to combat cancer.

1. Convolutional Neural Network(CNN)

Over the years, numerous studies have been conducted in an unwavering effort to improve early cancer detection and classification. Each of these investigations has made a distinct contribution to the ever-changing field of cancer treatment and research.

The inception of this endeavour occurred in 2018 with the publication of "Cancer Type Prediction Based on Chromatin 3D Structure and Copy Number Aberration with Convolutional Neural Networks" by Yuan et al. [85]. By employing CNN to represent intricate information such as chromatin 3D structure and copy number aberrations (CNA), the authors distinguished cancer patients from healthy individuals with an astounding 95% accuracy rate. This achievement highlights the considerable potential of computational oncology.

In 2019, Wen-Hui Chen et al. [86] investigated the combined use of pretreatment plasma Epstein-Barr virus (EBV) DNA levels and cervical node necrosis (CNN) as prognostic indicators for patients undergoing radiotherapy in regard to nasopharyngeal carcinoma (NPC), which is prevalent in Southeast Asia.The implementation of this integrated strategy unveiled an enhanced prognostic stratification method that exhibited potential in supporting treatment and management decisions for NPC. Mostavi et al. [87] introduced a methodology in their 2020 article "Convolutional Neural Network Models for Cancer Type Prediction via Gene Expression." Through ingenious representation of gene expression profiles as images, the authors enabled CNNs to extract crucial characteristics for accurate classification. Worth mentioning is the method's exceptional precision in differentiating cancer patients from individuals in good health.

2019 was a momentous year due to the publication of three pivotal studies. A method was proposed by Xia et al. [88] that utilised CNN to generate representations of DNA methylation profiles that resembled images. By combining the classification of cancer patients with five CNNs, their approach attained an impressive 95% accuracy rate, presenting a potentially fruitful avenue for augmenting the detection and diagnosis of cancer in its early stages. In their article, Hemlata et al. [89] tackled the complex issue of DNA sequence classification. Through the integration of SVMs and CNNs in hybrid models, their exhaustive analysis yielded exceptionally high classification accuracies on a wide variety of DNA sequences. This methodology possessed considerable promise in enhancing the precision and effectiveness of DNA sequence categorization, thereby creating avenues for more advanced biological investigations and implementations. Li et al. (2022) demonstrated the transformative capacity of machine learning (ML) in surface-enhanced Raman spectroscopy (SERS) within the domain of medical diagnostics in their article "Machine Learning using Convolutional Neural Networks for SERS Analysis of Biomarkers." The authors accomplished a noteworthy 95% accuracy in biomarker classification through the utilisation of CNNs to analyse SERS spectra of various biomarkers. This development holds the potential to enhance the sensitivity and specificity of medical diagnostics.

Zeng et al. [91] proposed an innovative method in their 2022 study "Serum Raman Spectroscopy Combined with Convolutional Neural Network for Rapid Diagnosis of HER2-Positive and Triple-Negative Breast Cancer" to address a critical issue in women's health and expedite the breast cancer diagnosis process. By shrewdly integrating serum Raman spectroscopy and CNNs, this approach successfully classified breast cancer subtypes with an impressive 91.11% accuracy. Wahid and Banday [92] introduced a methodology in their article "Classification of DNA Microarray Gene Expression Leukaemia Data via the ABC and CNN Method" in 2023. The authors of this study utilised CNNs for classification and an artificial bee colony (ABC) algorithm for feature extraction. As a result, they classified leukaemia data with an exceptional 98% accuracy, which bodes well for advances in early diagnosis and treatment.

The culmination of the exploration of these noteworthy papers was the research conducted by Tati Rajab Mengko et al. [93] in 2023, which was entitled "Join Classifier of Type and Index Mutation on Lung Cancer DNA Using Sequential Labelling Model." The authors proposed a novel methodology for the early detection of lung cancer by utilising a sequential labelling model and a CNN-powered pipeline to classify mutation types and indices within lung cancer DNA sequences. The outcomes of an assessment conducted on a collection of lung cancer DNA sequences were encouraging, as evidenced by an accuracy of 85.71% in identifying mutation indices and 82.35% in classifying mutation types; these results represent a substantial advancement in the field of cancer genomics.

Collectively, these investigations symbolise a logical sequence of steps taken to further the field of early cancer detection and classification; each study has made a significant contribution by introducing novel approaches and methodologies. Promising improved precision and effectiveness, the authors ultimately intend to transform the cancer care and research domain.

2. Recurrent Neural Network(RNN)

RNN assumes a critical function within the domain of DNA-based cancer detection. Particularly adept at capturing sequential dependencies within DNA sequences, these specialised DL models are indispensable for detecting subtle genetic variations linked to cancer. Advances in personalised healthcare and cancer research have been made possible in recent years by the combined efforts of computational techniques and genomics. In their article "Recurrent Neural Network for Genome Sequencing for Personalised Cancer Treatment in Precision Healthcare" [94] (2022), Gao et al. presented an innovative methodology. RNNs were utilised by the authors to accurately classify cancer types on the basis of genomic profiles. Capable of discerning enduring interdependencies within sequencing data, their approach attained an exceptional 99 percent precision, thereby offering hope for individualised cancer therapies guided by genetics.

Combining bidirectional LSTM RNNs and Rider-Chicken Optimisation (RCO), the method proposed by Aher et al. (2020) [95] successfully classifies cancer patients with a 95% degree of accuracy. This demonstrates the potential of computational methods in the context of cancer diagnosis using gene expression. In "Jaya Ant lion optimization-driven Deep recurrent neural network (DRNN) for cancer classification using gene expression data," Majji et al. (2021) [96] expanded upon this work. Utilising bidirectional LSTM Deep Recurrent Neural Networks (DRNN) optimised with Jaya Ant Lion Optimisation (JAO), they employ an innovative methodology. Its ability to classify cancer patients with an accuracy of 95.97% highlights the potential of DRNNs and

optimisation techniques to improve cancer classification through the utilisation of gene expression profiles. Together, these studies demonstrate the revolutionary potential of computational techniques in precision medicine and cancer research.

4. Other Deep Learning Methods

Recent years have witnessed notable progress at the intersection of DL and biological data analysis, which has the capacity to fundamentally alter our comprehension of intricate biological systems. Numerous noteworthy studies have surfaced, each making a contribution to the development of the field by integrating a variety of data sources and employing cutting-edge methodologies.

A multimodal deep learning approach was introduced in 2015 by Muxuan Liang [97], which utilised deep belief networks to analyse cancer data across multiple platforms. Their approach, which was not only groundbreaking but also exceptionally precise, surpassed conventional methods of classification. This research represented a turning point in the pursuit of more efficient methods for analysing cancer data. In the year 2017, Chen et al. [98] introduced an innovative methodology for the identification of genes associated with breast cancer. By integrating mRNA expression, DNA methylation, and copy number alterations into multi-omics data and utilising capsule networks to discover hierarchical gene representations, this methodology has the potential to revolutionise the way in which critical genes are identified in cancer research. A comprehensive survey was undertaken by Mahmud et al. [99] in 2018, with a particular focus on the profound possibilities that DL and reinforcement learning hold for the analysis of biological data. Deep reinforcement learning, a swiftly developing domain situated at the intersection of deep learning and reinforcement learning, was emphasised by the authors as being crucial for various tasks including drug discovery, medical image analysis, and protein structure prediction. The survey provided researchers with a significant asset as they navigated the dynamic realm of biological data analysis.

"MultiCapsNet" was unveiled in 2019 [100] as a deep learning model specifically engineered to resolve the persistent issue of data fusion by integrating heterogeneous data from variety of sources. By integrating the advantages of Capsule Networks and scCapsNets, this methodology improved interpretability by employing a routing mechanism that acquired knowledge of the significance of individual data sources. As a result, it made a valuable contribution to data integration endeavours in diverse fields, such as bioinformatics and medical imaging. A novel methodology utilising capsule networks, referred to as "i6mA-Caps," was presented in 2022 [101] to accurately detect DNA N6-methyladenine (6mA) sites. By capitalising on the capacity of capsule networks to acquire hierarchical data representations, this approach surpassed traditional machine learning methods in terms of precision and sensitivity, thereby expanding the frontiers of epigenetic site identification. In the year 2023, Mokoatle et al. (102) emphasised the importance of early cancer detection and suggested that sentence transformers be employed to represent DNA sequences in the classification of cancer. The results of their research demonstrated that XGBoost performed well with SimCSE representations, and further advancements could be made by implementing reinforcement networks. This provides insight into the potential of XGBoost in the field of cancer diagnosis in the future.

During the same calendar year, Jiaqi Li et al. [103] unveiled "DISMIR," an innovative methodology that utilises low-depth cell-free DNA sequencing to enhance the accuracy of cancer detection. By integrating DNA sequence and methylation data in an innovative manner and employing a DL model refined by Sentence BERT (SBERT) for enhanced performance and representation, this method provides a novel outlook on cancer detection. Mohammed et al. (104) re-introduced a hybrid cancer prediction model in 2023 [104]. This model incorporated reinforcement learning and multi-omics data. By optimising feature selection and ML classifier performance, this innovative approach holds promise for the identification of cancer biomarkers and the improvement of prediction accuracy. This represents a substantial advancement in the pursuit of more precise cancer prediction. The authors conclude their discussion of the applications of DL and reinforcement learning for the analysis of complex biological data in a 2023 paper by Mufti Mahmud et al. [105]. The authors emphasise the utilisation of DNA sequencing data in various applications, including medical image analysis, drug discovery, and genetic mutation identification. This extensive synopsis functions as a valuable reference for scientists seeking to apply these methodologies to practical biological dilemmas.

The aforementioned studies collectively illustrate a dynamic and ever-changing environment in which novel approaches to integrating data are expanding the limits of our capacity to analyse and comprehend intricate biological information. These advancements have the potential to have a substantial influence on cancer research

and diagnostics.

IV. Feature Selection:

1. Filter Methods

Within the field of cancer research, numerous studies have implemented novel methodologies with the aim of augmenting the precision and effectiveness of cancer categorization and biomarker detection. In the domain of DNA-based cancer detection, filter methods including the Chi-Square test, correlation-based feature selection, and Info Gain are utilised to discern pertinent genetic markers or characteristics that demonstrate noteworthy correlations with cancer. This aids in the prioritisation and selection of the most informative attributes, thereby ensuring precise classification.

Schwarzenbach et al. [106] investigated the potential of cell-free tumour DNA (cfDNA) as a marker for circulating tumour cells (CTCs) in the context of prostate cancer in 2009. The authors quantified cfDNA levels in blood samples from patients diagnosed with benign prostatic hyperplasia and prostate cancer using quantitative PCR. With significant correlations observed between cfDNA levels and the incidence of CTCs, the results were encouraging. Furthermore, it was observed that cfDNA and CTC levels were elevated in cases of metastasized prostate cancer, indicating their potential utility in the detection of metastatic disease and as a game-changer for early intervention.

Schwarzenbach et al. [107] shifted their attention to breast cancer in 2011. The purpose of their research was to assess the potential utility of cfDNA and RNA (cfRNA) as biomarkers. The levels of cfDNA and cfRNA were determined using quantitative PCR assays; breast cancer patients had substantially elevated levels compared to those with benign breast disease. Additional investigations were conducted into possible correlations among DNA concentrations, LOH frequency, RNA concentrations in blood serum, and clinical and histopathological risk factors among patients with breast cancer. These findings present promising opportunities for the management of breast cancer.

In 2013, Ammu Prasanna Kumar and Preeja Valsala [108] proposed a hybrid methodology that utilised high-dimensional DNA microarray data for feature selection in cancer classification. The method they implemented comprised phases of filtering and wrapping. For the wrapping phase, they utilised biogeography-based optimisation (BBO) and Pearson Product Moment Correlation (PPMC) in place of information gain (IG) and PPMC, respectively. By utilising back propagation neural networks and K nearest neighbour (KNN) evaluations, their method highlighted the importance of robust feature selection in cancer classification.

Al-Batah et al. [109] proposed an innovative methodology in 2019 that combined correlation-based feature selection and rules classifiers to efficiently discern relevant features within intricate gene microarray data with the aim of achieving accurate cancer classification. The approach demonstrated exceptional performance in comparison to modern methodologies, providing a robust instrument for precise cancer categorization within the domain of genomics investigation.

The year 2020 marked the advent of novel methodologies. In their seminal work, Kou et al. [110] unveiled a novel methodology that leverages DNA methylation data to discern cancer sample-specific associated genes (CSAGs), thereby providing a valuable instrument for differentiating cancerous samples from healthy ones. "G-Forest," a cost-sensitive feature selection method that integrates correlation-based feature selection and RFs, was introduced by Abdulla and Khasawneh [111]. This approach demonstrated superior classification accuracy compared to alternative techniques, all the while preserving a more limited subset of features.

Seo and Cho [112] introduced Dual Correlation Filters (DCFS), a feature selection algorithm designed specifically for the detection of cancer-associated somatic variants (CASVs) in high-dimensional genomic data, during the same year. Differentiating itself from conventional approaches, DCFS effectively mitigates overfitting by selecting features that are strongly correlated with the class label while minimising inter-feature correlations. By utilising this innovative method, CASV identification is significantly improved, leading to a more succinct CASV subset while maintaining classification accuracy.

Subsequently, in the same year, Alok Kumar Shukla et al. [113] proposed an innovative distributed correlation-based feature selection (D-CFS) approach to address the drawbacks of conventional CFS methods when examining noisy microarray data in high dimensions. D-CFS is a method that identifies significant biomarkers by distributing the data into smaller subgroups, performing CFS on each subset, and then aggregating the results.

2. Wrapper Methods

Innovative methods and techniques have been developed by scientists over the decades in an effort to enhance the detection and categorization of biomarkers associated with cancer. Wrapper methods are an essential category of feature selection techniques utilised in the domain of DNA-based cancer detection. The purpose of these methods is to assess the efficacy of machine learning models by analysing particular subsets of genes or genetic markers. This aids in the determination of the most informative DNA characteristics that are crucial for precise cancer diagnosis.

Manchester et al. [114] proposed an innovative method in 1988 for the identification of benzo[a]pyrene diol epoxide (BPDE)-DNA adducts in placental tissues from human subjects. The researchers employed Sequential Forward Selection (SFS), a highly efficient algorithm for feature selection, in order to discern informative features within the DNA methylation data. By establishing a correlation between BPDE-DNA adduct levels in placental samples and maternal benzo[a]pyrene exposure, this method provides a valuable instrument for comprehending the influence of environmental factors on the likelihood of developing cancer. In 1992, Alexandrov et al. [115] devised a fluorometric assay that was exceptionally sensitive and could quantify BPDE-DNA adducts in the lung tissue of individuals who were smokers. By surpassing its antecedents, this method enabled the examination of minute tissue samples and revealed notably elevated levels of adducts in individuals who smoke. The research emphasised the effectiveness of SFS in selecting features, thereby illuminating its capacity to improve the assessment of lung cancer risk and direct preventive measures.

DNA and protein adducts of Benzo[a]pyrene (BaP) in human tissues were investigated by Boysen and Hecht [116] in 2003 as potential biomarkers for BaP exposure and cancer risk. By utilising ELISA and IHC techniques, they were able to determine that cancer patients had substantially higher levels of adducts than non-cancerous patients. This highlights the importance of SFS in predictive modelling to identify critical features. This study demonstrated a substantial progression in comprehending the carcinogenesis induced by BaP and the potential of SFS in the identification of biomarkers. Hasri et al. [117] introduced a hybrid feature selection method that combines RFE and numerous SVMs for cancer classification in 2017. By surpassing alternative methodologies, this groundbreaking strategy improved the precision of classification on standard cancer datasets and presented prospects for utilisation in diagnostics and therapeutics.

In 2018, Ghosh et al. [118] introduced a comprehensive approach called "Genetic Algorithm based Cancerous Gene Identification (GAEF)" for analysing high-dimensional microarray data. This method utilised a GA and an ensemble of filter algorithms. The researchers effectively accounted for gene interactions in their approach, showcasing the efficacy of SFS in conjunction with Sequential Backward Selection (SBS) as a wrapper method to optimise feature selection for precise cancer gene identification. In 2020, Chen et al. [119] proposed a methodology for classifying breast cancer that improved SVM-Recursive Feature Elimination (SVM-RFE) by integrating LogFC normalisation, feature value equivalents, clustering analysis, and PCA. By surpassing current feature selection techniques, their method substantially enhances the accuracy of classification and presents encouraging potential for the development of advanced diagnostic instruments in the field of breast cancer research. Expanding upon the aforementioned advancements, Ghosh et al. [120] proposed a hybrid approach to feature selection in 2021, which aimed to identify cancer genes by merging an ensemble of filter methods with a GA. The application of SFS improved the performance of the model, demonstrating the potential of their methodology in the field of cancer research for biomarker discovery and classification accuracy. The cumulative endeavours described above symbolise an ongoing pursuit for enhanced efficacy and accuracy in the detection and categorization of cancer.

Table1: Summarizes the papers that detect other diseases apart from Cancer using DNA and ML.

Reference	Publication Year	Disease	Methods Used	Limitations	Accuracy
126	2021	Alzheimer's disease	Deep learning, embedded feature selection	Limited dataset, lack of external validation	92.20%
127	2022	Parkinson’s disease	AdaBoostClassifier algorithm	Dataset limited to people of the same age	89.72%.
128	2018	HIV	GLMNET, SVM, RF, and XGBoost.	The study was conducted on a single cohort of HIV-positive individuals and the results may not be generalizable to other populations.	95%
129	2019	Hepatitis B	XGBoost, RF, DCT, LR	Small datasets, requirement of	89.10%

				longitudinal follow-up data, and limited information considered	
130	2021	Tuberculosis	Machine learning algorithms, including C4.5, Random Forest, and Logitboost	The study was conducted on a relatively small sample size and the results may not be generalizable to other populations.	99%
131	2020	leukemia	Deep learning	Limited comparison to other state-of-the-art methods	75%
132	2020	Huntington's disease	decision tree, rule induction, random forest, and generalized linear model.	Data from one brain region, small sample size	97.46%
133	2022	Neisseria gonorrhoeae	Random Forest, Support Vector Machines (SVM), and Naive Bayes	The study is only focused on N. gonorrhoeae and that the models may not be generalizable to other organisms.	
134	2023	Monkeypox	LSTM and Gated Recurrent Unit.	Limited by the availability of data, the small number of patients	
135	2022	α-thalassemia	XGBoost, random forest and regularized logistic regression	Small sample size, limited to common α-thalassemia deletions	96%
136	2023	Glaucoma	XGBoost, random forest and regularized logistic regression	Small sample size, needs external validation	93%
137	2023	Multiple sclerosis	random forests, gradient boosting machines, and support vector machines.	Could not distinguish between primary progressive MS and secondary progressive MS	95%
138	2012	Alzheimer's disease	Partial Least Squares, Support Vector Machines	Small sample size, needs external validation	81%
139	2022	Parkinson's disease	Deep Belief Network (DBN)	Small sample size	95%
140	2019	Tuberculosis	Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM)	Small sample size	92.70%
141	2022	Monkeypox	Deep learning	Small dataset	82.96%
142	2023	Hepatitis B	Partial least squares-discriminant analysis (PLS-DA)	Small dataset	98.30%
143	2021	Thalassemia	Extreme gradient boosting (XGBoost)	Small sample size, needs external validation	93%
144	2020	COVID-19	Decision Tree, Discriminant Analysis, K-Nearest Neighbour, Support Vector Machine	Small dataset, needs external validation	98%

V. Other Machine Learning Techniques

Within the dynamic realm of biotechnology and medical research, significant progress has been achieved by scientists and researchers in the application of computational methodologies to augment our comprehension of complex biological systems and elevate healthcare results. Mukherjee's [121] research in 2000 centred on the utilisation of Shrinkage estimation and Support Vector Machines (SVMs), a computational methodology grounded in the principles of statistical learning theory, for the examination of DNA microarray data. Particularly when it comes to forecasting treatment outcomes for particular medical conditions, such as brain tumours and lymphoma, this methodology has demonstrated encouraging results. It is probable that the paper assesses the efficacy of SVM and other machine learning algorithms on a variety of molecular cancer datasets. By means of this research, the author makes a scholarly contribution to the expanding corpus of information concerning the effective and precise utilisation of sophisticated statistical techniques in analysing intricate genetic data. This undertaking is critical in order to extract significant biological insights from complex DNA microarray datasets. The utilisation of gene expression data for the classification of cancer is of utmost importance in the field of bioinformatics. The study conducted by Hong and Cho [122] in 2006 highlights the significance of employing ensemble methodologies to enhance the accuracy of DNA microarray data classification. The ensemble approach proposed by the authors is based on genetic programming and involves analysing the structure of classification criteria in order to assess diversity. Experiments conducted on widely used gene expression datasets demonstrate that this novel approach exhibits superior performance in comparison to conventional methods. Machine learning, as described by Sajda [123] in the same year, becomes a crucial instrument in developing complex, automated algorithms for biomedical data analysis. This review illuminates the most recent developments that

hold the potential to improve the detection, diagnosis, and monitoring of therapeutic interventions. The nuances of algorithmic construction, learning theory, and the inherent trade-offs in enhancing generalisation capabilities are emphasised. It provides examples of ML methodologies and their practical implementations in the field of biomedical diagnostics.

With the incidence of breast cancer on the rise, a systematic review conducted by Nassif et al. [124] in 2022 reveals the potential of artificial intelligence to improve detection and treatment. The potential of AI is significantly enhanced when early diagnosis is taken into account, as DL provides intricate insights into the various factors that impact disease detection. Although genetic-level analysis continues to be a costly approach, histopathological imaging has become the prevailing technique for detecting breast cancer. This study conducts a critical analysis of previous research and argues in favour of the utilisation of artificial intelligence in histopathological imaging and genetic sequencing as a means to diagnose and treat breast cancer. In 2023, Huang et al. [125] provide an exhaustive examination of the critical condition of lung cancer, which is characterised by its alarming rates of occurrence and death on a global scale. Despite the discouraging five-year survival rate as a whole, early diagnosis and prognosis are critical factors in achieving improved outcomes. This paper examines the revolutionary impact that artificial intelligence has had on lung cancer research, specifically in the areas of early detection and prognosis forecasting. The narrative skillfully incorporates a range of AI methodologies, such as natural language processing (NLP), ML, and DL, to emphasise their significant contributions to the clinical prognosis and diagnosis of lung cancer. Collectively, these studies provide an expansive perspective on the interplay between sophisticated computational approaches and the analysis of biological data, thereby laying the foundation for forthcoming advancements in the field of biomedicine.

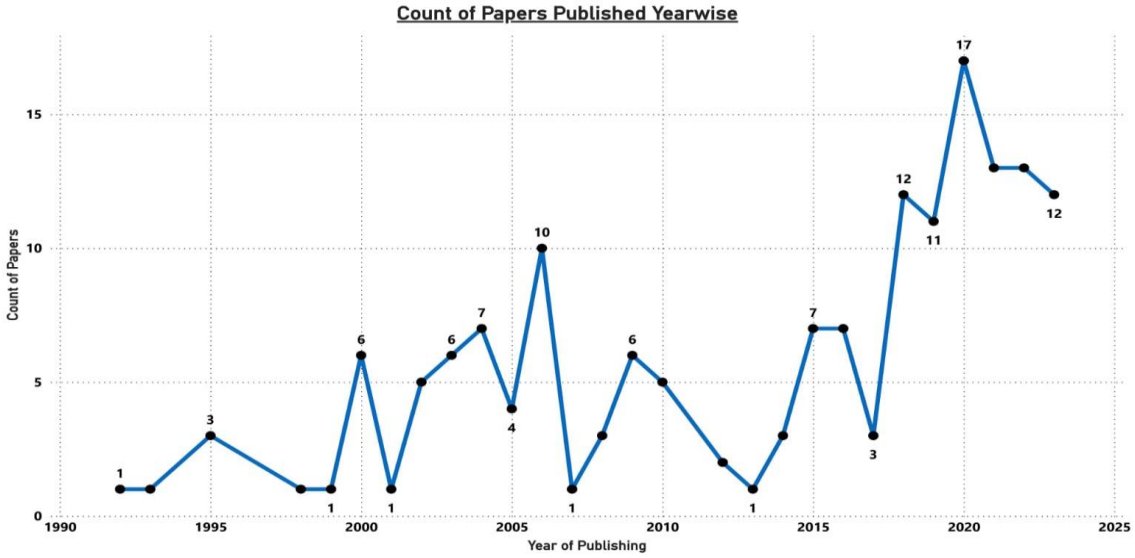

Fig1: This Line Graph shows that the number of papers published has increased steadily over time, with a significant surge in recent years.

2. Conclusion

This review explores the complex domain of cancer detection and classification, with a specific focus on DNA-based approaches that make use of ML and DL methodologies. It encompasses an extensive range of computational strategies, including advanced neural network architectures and dimensionality reduction techniques such as PCA and SVD. These methodologies have the capacity to fundamentally reshape the domain of cancer diagnostics. Although the assessment demonstrates notable advancements in accuracy and effectiveness, it is critical to recognise the utmost significance of DNA-based methodologies in the field of cancer detection. Frequently operating at the convergence of computational analysis and molecular biology, these methodologies are indispensable in elucidating the genetic indicators that signify early-stage cancer, thereby influencing the development of subsequent treatment protocols. The emphasises is on the continuous paradigm shift in oncological research through the utilisation of autoencoders, deep belief networks, and other novel DL models. This paradigm shift presents the potential for a future in which personalised cancer treatment is not only achievable through the utilisation of intricate insights derived from DNA analysis, but is also more precise.

3. Future Directions

As computational models continue to be increasingly integrated with clinical practice, the future of cancer detection and classification is positioned for transformative expansion. Potential for progress exists in the integration of ML methodologies with nascent biotechnological data, and in the implementation of these approaches on larger, more heterogeneous cohorts of patients. Additional investigation is necessary to authenticate these computational methodologies in practical contexts, surmount the obstacles posed by high-dimensional data, and guarantee the replicable nature of findings across disparate datasets. An imperative exists to create more resilient models capable of seamlessly integrating multi-omics data in order to provide a comprehensive perspective of the cancer domain. Furthermore, it is imperative to promote cooperation between computational scientists and clinicians to ensure that these technological developments can be effectively implemented in the clinic, with the ultimate goal of improving patient outcomes and survival rates. Investigating innovative computational methodologies, including advanced feature selection techniques and reinforcement learning, will significantly advance cancer research by paving the way for the next generation of breakthroughs that will elevate the discipline to the epicenter of precision medicine.

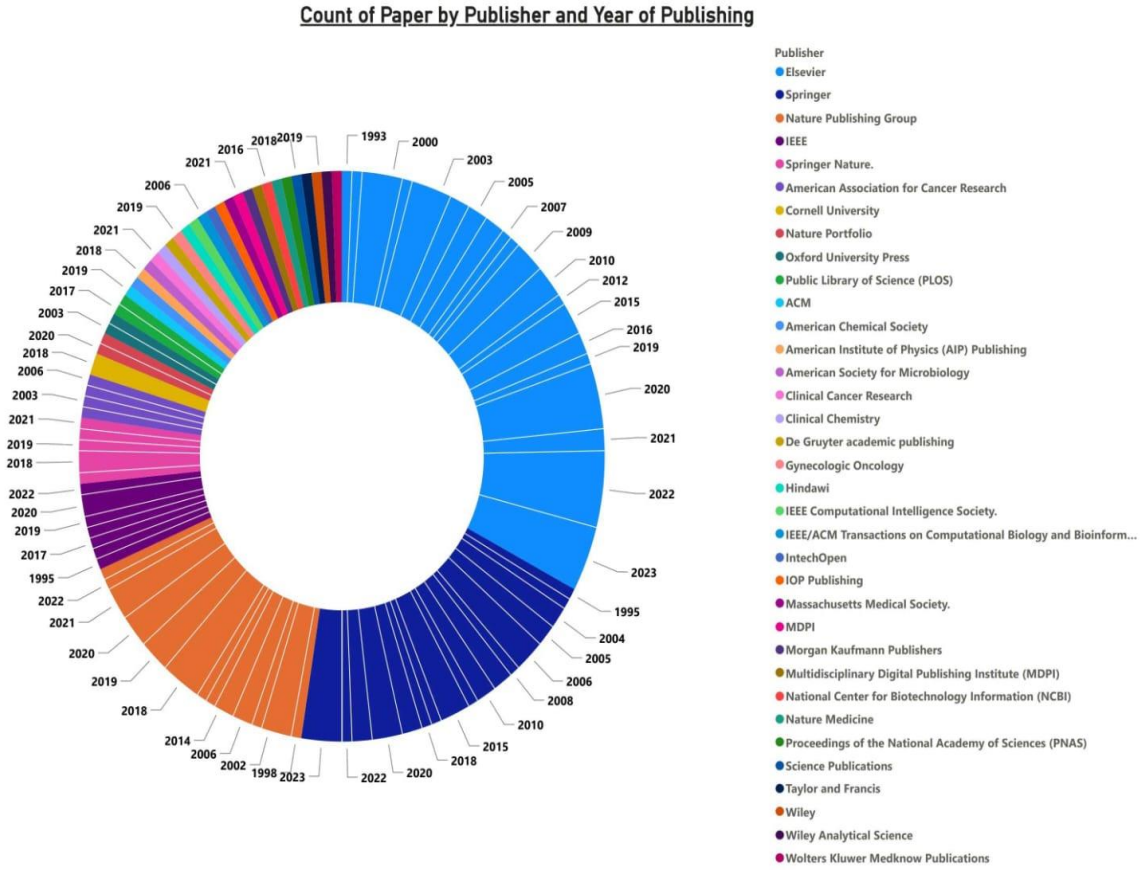

Fig2: This Donut Chart depicts the distribution of papers published between 1992 and 2023 categorized by publisher.

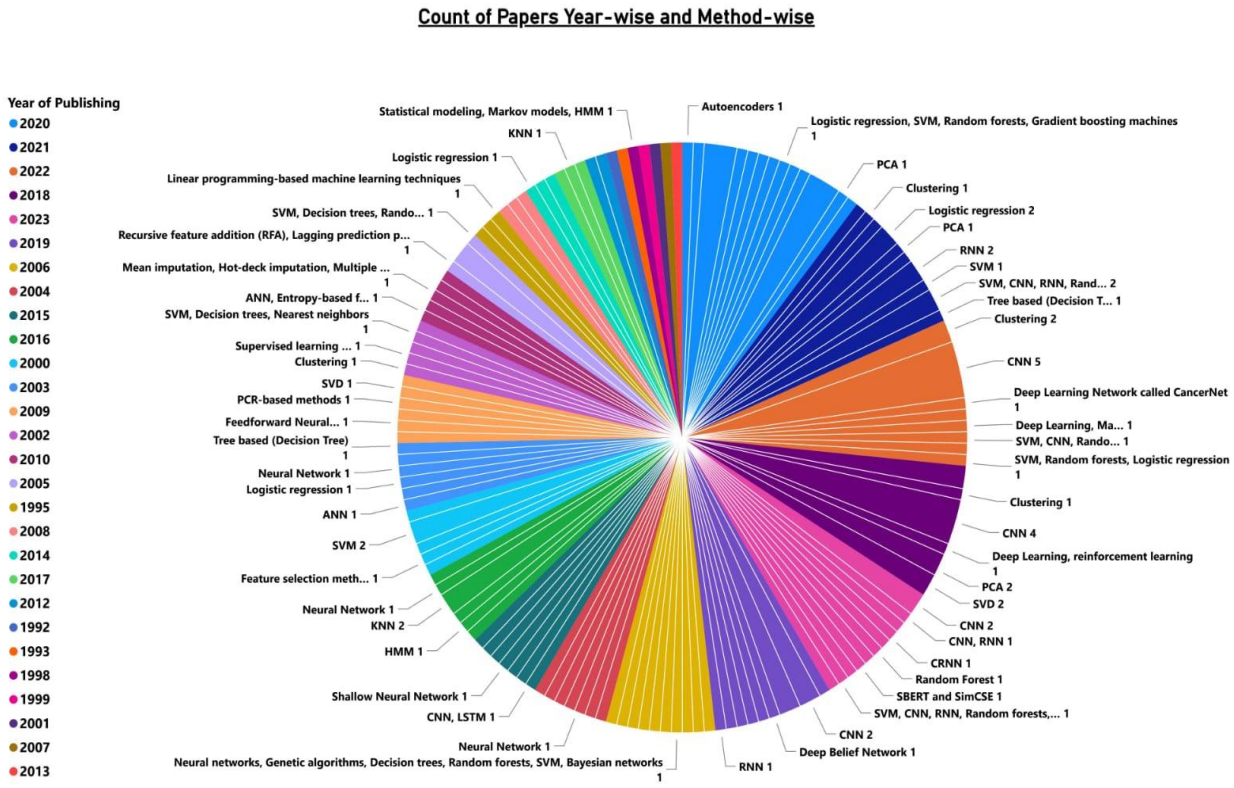

Fig3: This pie chart shows the distribution of different methods of ML and DL used in papers published between 1992 and 2023.

References

[1] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. Artificial intelligence in radiology. NatureReviews Cancer 2018;18(8):500–510. https://doi.org/10.1038/s41568-018-0016-5.

[2] Isaksson, M., Jalden, J., & Murphy, M. J. On using an adaptive neural network to predict lung tumor motion during respiration for radiotherapy applications. Medical Physics 2005;32(12):3801–3809. https://doi.org/10.1118/1.2134958.

[3] Brown MPS, GrundyWN, Lin D, Cristianini N, Sugnet C, Furey TS, et al. Knowledge-based analysis of microarray gene expression data using support vector machines. Proceedings of the National Academy of Sciences 2000;97(1):262–267. https://doi.org/10.1073/pnas.97.1.262.

[4] Guyon, Isabelle, Weston, Jason, Barnhill S, Vapnik, et al. Gene selection for cancer classification using support vector machines. Machine Learning 2002;46(1):389–407. https://doi.org/10.1023/A:1012487302797

[5] Valentini, Giorgio, Muselli, Marco, Ruffino F. Cancer recognition with bagged ensembles of support vector machines. Neurocomputing 2003;55(3):333–340. https://doi.org/10.1016/j.neucom.2003.09.001

[6] Guo S, Yan F, Xu J, et al. Identification and validation of the methylation biomarkers of non-small cell lung cancer (NSCLC). Clinical epigenetics 2015; 7(1):3. https://doi.org/10.1186/s13148-014-0035-3

[7] Bahado-Singh R, Vlachos KT, et al. Precision Oncology: Artificial Intelligence and DNA Methylation Analysis of Circulating Cell-Free DNA for Lung Cancer Detection. Frontiers in Oncology 2022;12. https://doi.org/10.3389/fonc.2022.790645

[8] Hosseini, M., Lotfi-Shahreza, M., & Nikpour, P. Integrative analysis of DNA methylation and gene expression through machine learning identifies stomach cancer diagnostic and prognostic biomarkers. Journal of Cellular and Molecular Medicine 2023; 27(5), 714-726. https://doi.org/10.1111/jcmm.17693.

[9] Venkatesan N, Pasupathy S, Gobinathan B. An efficient lung cancer detection using optimal SVM and improved

weight based beetle swarm optimization. Biomedical Signal Processing and Control 2023;p. 105373. https://doi.org/10.1016/j.bspc.2023.105373.

[10] Venkatesan N, Pasupathy S, Gobinathan B. An efficient lung cancer detection using optimal SVM and improved weight based beetle swarm optimization. Biomedical Signal Processing and Control 2023;p. 105373. https://doi.org/10.1016/j.bspc.2023.105373.

[11] Zhang N, Pasupathy S, Gobinathan B. An efficient lung cancer detection using optimal SVM and improved weight based beetle swarm optimization. Biomedical Signal Processing and Control 2023;67:105373. https://doi.org/10.1038/s41467-020-17316-z.

[12] Ahlquist DA, Zou H, Domanico M, et al. Next-generation stool DNA test accurately detects colorectal cancer and large adenomas. Gastroenterology. 2012;142(2):248-e26. https://doi.org/10.1053/j.gastro.2011.10.031.

[13] Sozzi G, Pastorino P, TagliabueM, Verderio C, Pizzamiglio L, Amadori G, et al. Quantification of Free Circulating DNA As a DiagnosticMarker in Lung Cancer. Journal of Clinical Oncology 2003;21(2):390–396. https://doi.org/10.1200/JCO.2003.02.006.

[14] HoqueMO, Begum S, Topaloglu O, Chatterjee A, Rosenbaum E, Van Criekinge W, et al. Quantitation of Promoter Methylation of Multiple Genes in Urine DNA and Bladder Cancer Detection. JNCI: Journal of the National Cancer Institute 2006 07;98(14):996–1004. https://doi.org/10.1093/jnci/djj265.

[15] Rasmussen SL, Krarup HB, Sunesen KG, Johansen MB, Stender MT, Pedersen IS, et al. Hypermethylated DNA, a circulating biomarker for colorectal cancer detection. PLoS ONE 2017;12(7):e0180809. https://doi.org/10.1371/journal.pone.0180809.

[16] Bach S, Paulis I, Sluiter NR, TibbesmaM,Martin I, van deWiel MA, et al. Detection of colorectal cancer in urine using DNA methylation analysis. Scientific Reports 2021;11(1):2363. https://doi.org/10.1038/s41598-021-81900-6.

[17] Wu TI, Huang RL, Su PH, Mao SP, Wu CH, Lai HC. Ovarian cancer detection by DNA methylation in cervical scrapings. Clinical Epigenetics 2019;11(1):166. https://doi.org/10.1186/s13148-019-0773-3.

[18] Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget Stool DNA Testing for Colorectal-Cancer Screening. The New England Journal of Medicine 2014;370(14):1287–1297. PMID: 24645800. https://doi.org/10.1056/NEJMoa1311194.

[19] Chalasani NP, Ramasubramanian TS, Bhattacharya A, Olson MC, Edwards V DK, Roberts LR, et al. A Novel Blood-Based Panel of Methylated DNA and Protein Markers for Detection of Early-Stage Hepatocellular Carcinoma. Clinical Gastroenterology and Hepatology 2021;19(12):2597–2605.e4. https://doi.org/10.1016/j.cgh.2020.08.065.

[20] Chen MC, Liao HC, Huang CL. Predicting Breast Tumor via Mining DNA Viruses with Decision Tree. In: 2006 IEEE International Conference on Systems, Man and Cybernetics, vol. 5; 2006. p. 3585–3589. https://doi.org/10.1109/ICSMC.2006.384685.

[21] Jerez-Aragonés JM, Gómez-Ruiz JA, Ramos-Jiménez G, Muñoz-Pérez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial Intelligence in Medicine 2003;27(1):45–63. https://doi.org/10.1016/S0933-3657(02)00086-6.

[22] Keerthika J, Sruthi D, Swathi D, Swetha S, Vinupriya R. Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1; 2021. p. 1530–1535. https://doi.org/10.1109/ICACCS51430.2021.9442043.

[23] Anglim PP, Galler JS, Koss MN, Hagen JA, Turla S, Campan M, et al. Identification of a panel of sensitive and specific DNA methylation markers for squamous cell lung cancer. Molecular Cancer 2008;7(1):62. https://doi.org/10.1186/1476-4598-7-62.

[24] Ponomaryova AA, Rykova EY, Cherdyntseva NV, Skvortsova TE, Dobrodeev AY, Zav'yalov AA, et al. Potentialities of aberrantly methylated circulating DNA for diagnostics and post-treatment follow-up of lung cancer patients. Lung Cancer;81(3):397–403. https://doi.org/10.1016/j.lungcan.2013.05.016.

[25] Cai Z, Xu D, Zhang Q, Zhang J, Ngai SM, Shao J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. Mol BioSyst 2015;11:791–800. https://doi.org/10.1039/c4mb00659c.

[26] Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics 2017;34(3):398–406. https://doi.org/10.1093/bioinformatics/btx622.

[27] Shen N, Du J, Zhou H, Chen N, Pan Y, Hoheisel JD, et al. A Diagnostic Panel of DNA Methylation Biomarkers for Lung Adenocarcinoma. Frontiers in Oncology 2019;9. https://doi.org/10.3389%2Ffonc.2019.01281.

[28] Chen Z, Xiong S, Li J, et al. DNA methylation markers that correlate with occult lymph node metastases of non-small cell lung cancer and a preliminary prediction model. Transl Lung Cancer Res. 2020;9(2):280-287. https://doi.org/10.21037/tlcr.2020.03.13.

[29] Wang Q, Zhou Y, Ding W, Zhang Z, Muhammad K, Cao Z. RandomForest with Self-Paced Bootstrap Learning in Lung Cancer Prognosis 2020;(1s). https://doi.org/10.1145/3345314.

[30] Aydadenta H, Adiwijaya A. A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest. Journal of Information Processing Systems 2018 Oct;14(5):1167–1175. https://doi.org/10.3745/JIPS.04.0087.

[31] Nuklianggraita T, Adiwijaya K, Aditsania A. On the Feature Selection of Microarray Data for Cancer Detection based on RandomForest Classifier. JURNAL INFOTEL 2020 08;12. http://dx.doi.org/10.20895/infotel.v12i3.485.

[32] Xie Y , Meng WY, Li RZ, Wang YW, Qian X, Chan C, et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. Translational Oncology 2021;14(1):100907. https://doi.org/10.1016/j.tranon.2020.100907.

[33] Li X, Yang T, Li S, Wang D, Song Y, Zhang S. Raman spectroscopy combined with principal component analysis and k nearest neighbour analysis for non-invasive detection of colon cancer. Laser Physics 2016 Feb;26(3):035702.

[34] Li Q, Hao C, Kang X, Zhang J, Sun X,WangW, et al. Colorectal Cancer and Colitis Diagnosis Using Fourier Transform Infrared Spectroscopy and an Improved K-Nearest-Neighbour Classifier. Sensors 2017 Nov;17(12):2739. https://doi.org/10.3390/s17122739.

[35] Kumar M, Rath NK, Rath SK. Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier. Journal of Biomedical Informatics 2016;60:395–409. https://doi.org/10.1016/j.jbi.2016.03.002.

[36] Bhuvaneswari P, Therese AB. Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm. Procedia Materials Science 2015;10:433–440. 2nd International Conference on Nanomaterials and Technologies (CNT 2014). https://doi.org/10.1016/j.mspro.2015.06.077.

[37] Zheng C, Xu R. Predicting cancer origins with a DNA methylation-based deep neural network model. PLoS ONE 2020;15(5):e0226461. https://doi.org/10.1371/journal.pone.0226461.

[38] Garro BA, Rodríguez K, Vázquez RA. Classification of DNA microarrays using artificial neural networks and ABC algorithm. Applied Soft Computing 2016;38:548–560. https://doi.org/10.1016/j.asoc.2015.10.002.

[39] Ringnér M, Peterson C. Microarray-based cancer diagnosis with artificial neural networks. BioTechniques 2003;34:S30–S35. https://doi.org/10.2144/mar03ringner.

[40] Spahic L, Cordic S. Prostate Tissue Classification Based on Prostate-Specific Antigen Levels and Mitochondrial DNA Copy Number Using Artificial Neural Network. In: Badnjevic A, Škrbic R, Gurbeta Pokvic L, editors. CMBEBIH 2019 Cham: Springer International Publishing; 2020. p. 649–654. https://doi.org/10.1007/978-3-030-17971-7_98.

[41] Marchevsky AM, Tsou JA, Laird-Offringa IA. Classification of Individual Lung Cancer Cell Lines Based on DNA Methylation Markers: Use of Linear Discriminant Analysis and Artificial Neural Networks. The Journal of Molecular Diagnostics 2004;6(1):28–36. https://doi.org/10.1016/S1525-1578(10)60488-6.

[42] Cahyaningrum K, Adiwijaya, Astuti W. Microarray Gene Expression Classification for Cancer Detection using Artificial Neural Networks and Genetic Algorithm Hybrid Intelligence. In: 2020 International Conference on Data Science and Its Applications (ICoDSA); 2020. p. 1–7. https://doi.org/10.1109/ICoDSA50139.2020.9213051.

[43] Atlam M, Torkey H, Salem H, El-Fishawy N. A New Feature Selection Method for Enhancing Cancer Diagnosis Based on DNA Microarray. In: 2020 37th National Radio Science Conference (NRSC); 2020. p. 285–295. https://doi.org/10.1109/NRSC49500.2020.9235095.

[44] Shahbaz M, Faruq S, Shaheen M, Masood SA. Cancer Diagnosis Using Data Mining Technology. Life Science Journal 2012;9(1):308–313.

[45] Maliha SK, Ema RR, Ghosh SK, Ahmed H, Mollick MRJ, Islam T. Cancer Disease Prediction Using Naive Bayes, K-Nearest Neighbor and J48 algorithm. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 2019. p. 1–7. https://doi.org/10.1109/ICCCNT45670.2019.8944686

[46] Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 2006 07;22(14):e184–e190. https://doi.org/10.1093/bioinformatics/btl230.

[47] Rouprêt M, Hupertan V, Yates DR, Comperat E, Catto JWF, Meuth M, et al. A comparison of the performance of microsatellite and methylation urine analysis for predicting the recurrence of urothelial cell carcinoma, and definition of a set of markers by Bayesian network analysis. British Journal of Cancer 2008;98(5):873–879. https://doi.org/10.1111/j.1464-410x.2008.07591.x.

[48] Oni´sko A, Druzdzel MJ, Austin RM. Application of Dynamic Bayesian Networks to Cervical Cancer Screening. Intelligent Information Systems 2009;9999:1–10.

[49] Kourou K, Rigas G, Papaloukas C, Mitsis M, Fotiadis DI. Cancer classification from time series microarray data through regulatory Dynamic Bayesian Networks. Computers in Biology and Medicine 2020;116:103577. https://doi.org/10.1016/j.compbiomed.2019.103577

[50] Yu X, Sun S. HMM-DM: identifying differentially methylated regions using a hidden Markov model. Statistical Applications in Genetics and Molecular Biology 2016;15(1):69–81. https://doi.org/10.1515/sagmb-2015-0077.

[51] Momenzadeh M, Sehhati M, Rabbani H. Using hidden Markov model to predict recurrence of breast cancer based on

sequential patterns in gene expression profiles. Journal of Biomedical Informatics 2020;111:103570. https://doi.org/10.1016/j.jbi.2020.103570.

[52] Friedrich D, Jin C, Zhang Y , Demin C, Yuan L, Berynskyy L, et al. In: Tolxdorff T, Deserno TM, Handels H,Meinzer HP, editors. Identification of Prostate Cancer Cell Nuclei for DNA-Grading Of Malignancy Springer Berlin Heidelberg; 2012. p. 334–339. https://doi.org/10.1007/978-3-642-28502-8_58.

[53] Bhola A, Tiwari A. Machine Learning Based Approaches for Cancer Classification Using Gene Expression Data. Machine Learning and Applications: An International Journal 2015 12;2:01–12. http://dx.doi.org/10.5121/mlaij.2015.2401.

[54] Hajianfar G, Shiri I, Maleki H, Oveisi N, Haghparast A, Taghipour A. Noninvasive O6-methylguanine-DNA methyltransferase status prediction in glioblastoma multiforme cancer using magnetic resonance imaging radiomics features: univariate and multivariate analysis. World Neurosurgery 2019;132:e140-e161. https://doi.org/10.1016/j.wneu.2019.08.232.

[55] Nuhic J, Spahic L, Cordic S, Kevric J. Comparative Study on Different Classification Techniques for Ovarian Cancer Detection. In: Badnjevic A, Škrbic R, Gurbeta Pokvic L, editors. CMBEBIH 2019 Cham: Springer International Publishing; 2020. p. 511–518. https://doi.org/10.1007/978-3-030-17971-7_76.

[56] Uthamacumaran A, Elouatik S, AbdouhM, Berteau-Rainville M, Gao Z, Arena G. Machine learning characterization of cancer patients-derived extracellular vesicles using vibrational spectroscopies: results from a pilot study. Applied Intelligence 2022;52(11):12737–12753. https://doi.org/10.1007/s10489-022-03203-1.

[57] Yu D, Liu Z, Su C, Han Y, Duan X, Zhang R, et al. Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier. Thorac Cancer. 2020;11(1):95-102. https://doi.org/10.1111/1759-7714.13204.

[58] Atlam M, Torkey H, Salem H, El-Fishawy N. A new feature selection method for enhancing cancer diagnosis based on DNA microarray. 2020 37th National Radio Science Conference (NRSC) 2020;p. 285–295. https://doi.org/10.1109/NRSC49500.2020.9235095.

[59] Swathi K, Kodukula S. XGBoost Classifier with Hyperband Optimization for Cancer Prediction Based on Gene selection by Using Machine Learning Techniques. Revue d'Intelligence Artificielle 2022;36(5):655–670. https://doi.org/10.18280/ria.360502.

[60] Ma B, Meng F, Yan G, Yan H, Chai B, Song F. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in Biology and Medicine 2020;121:103761. https://doi.org/10.1016/j.compbiomed.2020.103761.

[61] Liu H, Qiu C, Wang B, Bing P, Tian G, et al. Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. Frontiers in Cell and Development Biology 2021;9: 619330. https://doi.org/10.3389/fcell.2021.619330.

[62] Banjar A, Ali F, Alghushairy O, Daud A. iDBP-PBMD: A machine learning model for detection of DNA-binding proteins by extending compression techniques into evolutionary profile. Chemometrics and Intelligent Laboratory Systems 2022;231:104697. https://doi.org/10.1016/j.chemolab.2022.104697.

[63] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, AngeloM, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences 2001;98(26):15149–15154. https://doi.org/10.1073/pnas.211566398.

[64] Amor, R.d., Colomer, A., Monteagudo, C. et al. A deep embedded refined clustering approach for breast cancer distinction based on DNA methylation. Neural Computing and Applications 2022;34(13):10243 – 10255. https://doi.org/10.1007/s00521-021-06357-0.

[65] Li LW, Xiaosheng. Identification of gastric cancer subtypes based on pathway clustering. npj Precision Oncology 2021;5(1):1 – 46. https://doi.org/10.1038/s41698-021-00186-z.

[66] Virmani AK, Tsou JA, Siegmund KD, Shen LY, Long TI, Laird PW, et al. Hierarchical clustering of lung cancer cell lines using DNA methylation markers. Cancer epidemiology, biomarkers & prevention 2002;11(3):291–297.

[67] Gollapalle E,Wang R, Adetolu R, TsaoD, FranciscoD, Sigounas G, et al. Detection of Oxidative Clustered DNA Lesions in X-Irradiated Mouse Skin Tissues and Human MCF-7 Breast Cancer Cells. Radiat Res 2007;167:207–216. https://doi.org/10.1667/rr0659.1.

[68] Gunasekaran M, Vijayakumar V, Varatharajan R, M K P, Sundarasekar R, Hsu CH. Machine Learning Based Big Data Processing Framework for Cancer Diagnosis Using Hidden Markov Model and GM Clustering. Wireless Personal Communications 2018; 10;102. https://doi.org/10.1007/s11277-017-5044-z.

[69] Sehhati M, Tabatabaiefar MA, Gholami AH, Sattari M. Using Classification and K-means Methods to Predict Breast Cancer Recurrence in Gene Expression Data. Journal of Medical Signals and Sensors 2022;12(1):122 – 126. https://doi.org/10.4103%2Fjmss.jmss_117_21.

[70] Adiwijaya, Wisesty UN, Lisnawati E, Aditsania A, Kusumo DS. Dimensionality Reduction using Principal Component

Analysis for Cancer Detection based on Microarray Data Classification. Journal of Computer Science 2018 Nov;14(11):1521–1530. https://doi.org/10.3844/jcssp.2018.1521.1530.

[71] Menaga D, Revathi S. Probabilistic Principal Component Analysis (PPCA) Based Dimensionality Reduction and Deep Learning for Cancer Classification. In: Dash SS, Das S, Panigrahi BK, editors. Intelligent Computing and Applications Springer Singapore; 2021. p. 353–368. https://doi.org/10.1007/978-981-15-5566-4_31.

[72] Agarwal A, Venkatakrishna K, Tan B. Cellular DNA based cancer diagnosis using self-internalized plasmonic sensors. Sensors and Actuators B: Chemical 2020;321:128496. https://doi.org/10.1016/j.snb.2020.128496.

[73] Harris AT, Lungari A, Needham CJ, Smith SL, Lones MA, Fisher SE, et al. Potential for Raman spectroscopy to provide cancer screening using a peripheral blood sample. Head and Neck Oncology 2009;1(1):34. https://doi.org/10.1186/1758-3284-1-34.

[74] Feng S, Lin J, Huang Z, Chen G, Chen W, Wang Y, et al. Esophageal cancer detection based on tissue surface-enhanced Raman spectroscopy and multivariate analysis. Journal of Biomedical Optics 2018;102(4):043702. https://doi.org/10.1063/1.4789996.

[75] Huynh HT, Kim JJ, Won Y. Classification Study on DNA Microarray with Feedforward Neural Network Trained by Singular Value Decomposition. International Journal of Bio-Science and Bio-Technology 2009 12;1(1):17–24.

[76] Doufekas K, Zheng SC, Ghazali S, Wong M, Mohamed Y, Jones A, et al. DNA Methylation Signatures in Vaginal Fluid Samples for Detection of Cervical and Endometrial Cancer. International Journal of Gynecological Cancer 2016 7;26(7):1384–1391. https://doi.org/10.1097/igc.0000000000000739.

[77] Jiang Y, Wang R, Urrutia E, Anastopoulos IN, Nathanson KL, Zhang NR. CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. Genome Biology 2018 11;19(1):202. https://doi.org/10.1186/s13059-018-1578-y.

[78] Al-Juniad AF, Qaid TS, Al-Shamri MYH, Ahmed MHA, Raweh AA. Vertical and Horizontal DNA Differential Methylation Analysis for Predicting Breast Cancer. IEEE Access 2018;6:53533–53545. https://doi.org/10.1109/ACCESS.2018.2871027.

[79] Franco EF, Rana P, Cruz A, Calderón VV, Azevedo V, Ramos RTJ, et al. Performance Comparison of Deep Learning Autoencoders for Cancer Subtype Detection Using Multi-Omics Data. Cancers 2021 Apr;13(9):2013. https://doi.org/10.3390/cancers13092013.

[80] Khwaja M, KalofonouM, Toumazou C. A Deep Autoencoder System for Differentiation of Cancer Types Based on DNA Methylation State; 2018. https://doi.org/10.48550/arXiv.1810.01243.

[81] Chen X, Xie J, Yuan Q. A Method to Facilitate Cancer Detection and Type Classification from Gene Expression Data using a Deep Autoencoder and Neural Network; 2018. https://doi.org/10.48550/arXiv.1812.08674.

[82] Zhang X, Zhang J, Sun K, Yang X, Dai C, Guo Y. Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2019. p. 765–769. https://doi.org/10.1109/BIBM47256.2019.8983228.

[83] Adem K. Diagnosis of breast cancer with Stacked autoencoder and Subspace kNN. Physica A: Statistical Mechanics and its Applications 2020;551:124591. https://doi.org/10.1016/j.physa.2020.124591.

[84] Wisesty UN, Pratama BPB, Aditsania A, Adiwijaya. Cancer Detection Based on Microarray Data Classification Using Deep Belief Network and Mutual Information. In: 2017 5th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME); 2017. p. 157–162. https://doi.org/10.1109/ICICI-BME.2017.8537736.

[85] Ronoud S, Asadi S. An evolutionary deep belief network extreme learning-based for breast cancer diagnosis. Soft Computing 2019;23(24):13139–13159. https://doi.org/10.1007/s00500-019-03856-0.

[86] Yuan Y, Shi Y, Su X, Zou X, Luo Q, Feng DG, et al. Cancer type prediction based on copy number aberration and chromatin 3D structure with convolutional neural networks. BMC Genomics 2018;19(Suppl 6):S565. https://doi.org/10.1186/s12864-018-4919-z.

[87] Chen WH, Du YY, Luo XS, Tang LQ, Mai HQ, Chen QY, et al. Combining pretreatment plasma Epstein-Barr virus DNA level and cervical node necrosis improves prognostic stratification in patients with nasopharyngeal carcinoma: A cohort study. Cancer Med 2019;8(16):6841–6852. https://doi.org/10.1002/cam4.2481.

[88] Mostavi M, Chiu YC, Huang Y, Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. BMC Medical Genomics 2020;16:1–13. https://doi.org/10.1186/s12920-020-0677-2.

[89] Xia C, Xiao Y, Wu J, Zhao X, Li H. A Convolutional Neural Network Based Ensemble Method for Cancer Prediction Using DNA Methylation Data; 2019. p. 191–196. https://doi.org/10.1145/3318299.3318372.

[90] Gunasekaran, Hemalatha, Ramalakshmi K, Arokiaraj A, Macedo R, Deepa Kanmani S, et al. Analysis of DNA Sequence Classification Using CNN and Hybrid Models. Computational and Mathematical Methods in Medicine 2022;2021:1–12. https://doi.org/10.1155/2021/1835056.

[91] Li Qiaoyi J, Dukes PV, LeeW, SarkisM, Vo-Dinh T. Machine learning using convolutional neural networks for SERS analysis of biomarkers in medical diagnostics. Journal of Raman Spectroscopy 2022;53(6):797–805. https://doi.org/10.1002/jrs.6447.

[92] Zeng Q, Chen C, Chen C, Song H, Li M, Yan J, et al. Serum Raman spectroscopy combined with convolutional neural network for rapid diagnosis of HER2-positive and triple-negative breast cancer. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2023;286:122000. https://doi.org/10.1016/j.saa.2022.122000.

[93] Wahid A, Banday MT. Classification of DNA Microarray Gene Expression Leukaemia Data through ABC and CNN Method. International Journal of Intelligent Systems and Applications in Engineering 2023 Jul;11(7s):119–131.

[94] Wisesty UN, Purwarianti A, Pancoro A, Chattopadhyay A, Phan NN, Chuang EY, et al. Join Classifier of Type and Index Mutation on Lung Cancer DNA Using Sequential Labeling Model. IEEE Access 2022;10:9004–9021. https://doi.org/10.1109/ACCESS.2022.3142925.

[95] Suresh, A., Nair, R.R., Neeba, E.A. et al. RETRACTED ARTICLE: Recurrent Neural Network for Genome Sequencing for Personalized Cancer Treatment in Precision Healthcare. Neural Process Lett 2023; 55, 875.https://doi.org/10.1007/s11063-021-10572-0.

[96] Aher CN & Jena AK. Rider-chicken optimization dependent recurrent neural network for cancer detection and classification using gene expression data, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2021; 9(2): 174-191. https://doi.org/10.1080/21681163.2020.1830436.

[97] Majji R, Nalinipriya G, Vidyadhari C. et al. Jaya Antlion optimization-driven Deep recurrent neural network for cancer classification using gene expression data. Medical Biol Eng Comput 2021;59(5):1005–1021. https://doi.org/10.1007/s11517-021-02350-w.

[98] Liang M, Li Z, Chen T, Zeng J. Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2015;12(4):928–937. https://doi.org/10.1109/TCBB.2014.2377729.

[99] Peng C, Zheng Y, Huang DS. Capsule Network Based Modeling of Multi-omics Data for Discovery of Breast Cancer-Related Genes. IEEE/ACMTransactions on Computational Biology and Bioinformatics 2020;17(5):1605–1612. https://doi.org/10.1109/TCBB.2019.2909905.

[100] Liu M, Shen X, Pan W. Deep reinforcement learning for personalized treatment recommendation. Statistics in Medicine 2022 06;41. https://doi.org/10.1002/sim.9491.

[101] Wang L, Miao X, Zhang J, Cai J. Multi CapsNet: a interpretable deep learning classifier integrate data from multiple sources. bioRxiv 2019;. https://doi.org/10.1101/570507.

[102] Rehman MU, Tayara H, Zou Q, Chong KT. i6mA-Caps: a CapsuleNet-based framework for identifying DNA N6-methyladenine sites. Bioinformatics 2022 06;38(16):3885–3891. https://doi.org/10.1093/bioinformatics/btac434.

[103] Mokoatle M, Marivate V, Mapiye D, Bornman R, Hayes VM. A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application. BMC Bioinformatics 2023;24(112):1–21. https://doi.org/10.1186/s12859-023-05235-x.

[104] Li J, Wei L, Zhang X, Zhang W, Wang H, Zhong B, et al. DISMIR: Deep learning-based noninvasive cancer detection by integrating DNA sequence and methylation information of individual cell-free DNA reads. https://doi.org/10.1093/bib/bbab250.

[105] Mohammed MA, Lakhan A, Abdulkareem KH, Garcia-Zapirain B. A hybrid cancer prediction based on multi-omics data and reinforcement learning state action reward state action (SARSA). Computers in Biology and Medicine 2023;154:106617. https://doi.org/10.1016/j.compbiomed.2023.106617.

[106] Mahmud M, Kaiser MS, Hussain A, Vassanelli S. Applications of Deep Learning and Reinforcement Learning to Biological Data. IEEE Transactions on Neural Networks and Learning Systems 2018;29(6):2063–2079. https://doi.org/10.1109/TNNLS.2018.2790388.

[107] Schwarzenbach H, Alix-Panabieres C, Muller I, Letang N, Vendrell JP, Rebillard X, et al. Cell-free Tumor DNA in Blood Plasma As a Marker for Circulating Tumor Cells in Prostate

[108] Cancer. Clinical Cancer Research 2009 02;15(3):1032–1038. https://doi.org/10.1158/1078-0432.CCR-08-1910.

[109] Schwarzenbach H, Muller V, Milde-Langosch K, Steinbach B, Pantel K. Evaluation of cell-free tumour DNA and RNA in patients with breast cancer and benign breast disease. Mol BioSyst 2011;7:2848–2854. https://doi.org/10.1039/C1MB05197K.

[110] Kumar AP, Valsala P. Feature Selection for High Dimensional DNA Microarray Data Using Hybrid

[111] Approaches. Bioinformation 2013;9(16):824–828. https://doi.org/10.6026%2F97320630009824.

[112] Al-Batah MS, Zaqaibeh BM, Alomari SA, Alzboon MS. Gene Microarray Cancer Classification using Correlation Based Feature Selection Algorithm and Rules Classifiers. International Journal of Online and Biomedical Engineering (iJOE) 2019May;15(08):pp. 62–73. https://doi.org/10.3991/ijoe.v15i08.10617.

[113] Kou C, Zhang Y, Wang J. CSSIG: Identification of Cancer Sample-Specific Associated Genes Using Information Gain Based on DNA Methylation Data. ICBDT '20; 2020. p. 136–140 https://doi.org/10.1145/3422713.3422740.

[114] Abdulla M, Khasawneh MT. G-Forest: An ensemble method for cost-sensitive feature selection in gene expression microarrays. Artificial Intelligence in Medicine 2020;108:101941. https://doi.org/10.1016/j.artmed.2020.101941.

[115] Seo H, Cho DH. Feature selection algorithm based on dual correlation filters for cancer-associated somatic variants. BMC Bioinformatics 2020 10;21(1):486. https://doi.org/10.1186/s12859-020-03767-0.

[116] Shukla AK, Tripathi D. Detecting biomarkers from microarray data using distributed correlation based gene selection. Genes & Genomics 2020 04;42(4):449–465. https://doi.org/10.1007/s13258-020-00916-w.

[117] Boysen, Gunnar, and Stephen S Hecht. Analysis of DNA and protein adducts of benzo[a]pyrene in human tissues using structure-specific methods. Mutation research 2003 vol. 543,1: 17-30. https://doi.org/10.1016/s1383-5742(02)00068-6.

[118] Alexandrov K, Rojas M, Geneste O, Castegnaro M, Camus AM, Petruzzelli S, et al. An Improved Fluorometric Assay for Dosimetry of Benzo(a)pyrene Diol-Epoxide-DNA Adducts in Smokers' Lung: Comparisons with Total Bulky Adducts and Aryl Hydrocarbon Hydroxylase Activity. Cancer Research

[119] 1992 11;52(22):6248–6253.[120] Boysen G, Hecht SS. Analysis of DNA and protein adducts of benzo[a]pyrene in human tissues using structure-specific methods. Mutation Research/Reviews in Mutation Research 2003;543(1):17–30. https://doi.org/10.1016/s1383-5742(02)00068-6.

[121] Hasri NM, Wen NH, Howe CW, Mohamad MS, Deris S, Kasim S. Improved support vector machine using multiple SVM-RFE for cancer classification. International Journal on Advanced Science, Engineering and Information Technology 2017;7(4-2):1589–1594.

[122] Ghosh, M., Adhikary, S., Ghosh, K.K. et al. Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Med Biol Eng Comput 2019; 57: 159–176. https://doi.org/10.1007/s11517-018-1874-4.

[123] Chen G, Xie X, Li S. Research on Complex Classification Algorithm of Breast Cancer Chip Based on SVM-RFE Gene Feature Screening. Complexity 2020;2020:1342874. https://doi.org/10.1155/2020/1342874.

[124] Kanti Ghosh K, Begum S, Sardar A, Adhikary S, Ghosh M, Kumar M, et al. Theoretical and empirical analysis of filter ranking methods: Experimental study on benchmark DNA microarray data. Expert Systems with Applications 2021;169:114485. https://doi.org/10.1016/j.eswa.2020.114485.

[125] Mukherjee S. Application of statistical learning theory to DNA microarray analysis. PhD thesis, Citeseer; 2000.

[126] Hong JH, Cho SB. The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. Artificial intelligence in Medicine 2006;36(1):43–58. https://doi.org/10.1016/j.semcancer.2023.01.006.

[127] Sajda P. Machine learning for detection and diagnosis of disease. Annu Rev Biomed Eng 2006;8:537–565. https://doi.org/10.1146/annurev.bioeng.8.061505.095802.

[128] Nassif AB, Talib MA, Nasir Q, Afadar Y, Elgendy O. Breast cancer detection using artificial intelligence techniques: A systematic literature review. Artificial Intelligence in Medicine 2022;127:102276. https://doi.org/10.1016/j.artmed.2022.102276.

[129] Huang S, Yang J, ShenN,XuQ, ZhaoQ. Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. In: Seminars in Cancer Biology Elsevier; 2023. https://doi.org/10.1016/j.artmed.2005.06.002.

[130] Mahendran N, P M DRV. A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. Computers in Biology and Medicine 2022;141:105056. https://doi.org/10.1016/j.compbiomed.2021.105056.

[131] Makarious MB, Leonard HL, Vitale D, Iwaki H, Sargent L, Dadu A, et al. Multi-modality machine learning predicting Parkinson's disease. npj Parkinson's Disease 2022;8(1):35. https://doi.org/10.1038/s41531-022-00288-w.

[132] Zhang X, Hu Y, Aouizerat BE, Peng G, Marconi VC, Corley MJ, et al. Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality. Clinical Epigenetics 2018;10(1):155. https://doi.org/10.1186/s13148-018-0591-z.

[133] Chuzhanova NA, Tian X, Chong Y, Huang Y, Guo P, Li M, et al. Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance. Computational and Mathematical Methods in Medicine 2019;2019:6915850. https://doi.org/10.1155/2019/6915850.

[134] Hadikurniawati W, Anwar MT, Marlina D, Kusumo H. Predicting tuberculosis drug resistance using machine learning based on DNA sequencing data. Journal of Physics Conference Series 2021;1869(1):012093.

https://doi.org/10.1088/1742-6596/1869/1/012093.

[135] He J, Pu X, Li M, Li C, Guo Y. Deep convolutional neural networks for predicting leukemia-related transcription factor binding sites fromDNA sequence data. Chemometrics and Intelligent Laboratory Systems 2020;199:103976. https://www.sciencedirect.com/science/article/pii/S0169743919306367.

[136] Cheng J, Liu HP, Lin WY, Tsai FJ. Identification of contributing genes of Huntington's disease by machine learning. BMC Medical Genomics 2020;13(1):176. https://doi.org/10.1186/s12920-020-00822-w.

[137] Martin SL, Mortimer TD, Grad YH. Machine learning models for Neisseria gonorrhoeae antimicrobial susceptibility tests. Annals of the New York Academy of Sciences 2023;1520(1):74–88. https://doi.org/10.1111/nyas.14549.

[138] Pathan RK, UddinMA, Paul AM, Uddin MI, Hamd ZY, Aljuaid H, et al. Monkeypox genome mutation analysis using a time series model based on long short-termmemory. PLOS ONE 2023 08;18(8):1–17. https://doi.org/10.1371/journal.pone.0290045.

[139] Doan PL, Nguyen DA, Le QT, Hoang DTT, Nguyen HD, Nguyen CC, et al. Detection of maternal carriers of common thalassemia deletions from cell-free DNA. Scientific Reports 2022;12(1):13581. https://doi.org/10.1038/s41598-022-17718-7.

[140] Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. PLOS ONE 201705;12(5):1–16. https://doi.org/10.1371/journal.pone.0177726.

[141] Barbour C, Kosa P, KomoriM, TanigawaM,Masvekar R,Wu T, et al. Molecular-based diagnosis of multiple sclerosis and its progressive stage. Annals of Neurology 2017;82(5):795–812.

[142] https://doi.org/10.1002/ana.25083.

[143] Segovia F, Górriz JM, Ramírez J, Salas-González D, Álvarez I. Early diagnosis of Alzheimer's disease based on Partial Least Squares and Support Vector Machine. Expert Systems with Applications 2013;40(2):677–683. https://doi.org/10.1016/j.eswa.2012.07.071.

[144] Ibrahim, Ahmed Zohair, Prakash P, Sakthivel V, Prabu P. Integrated Approach of Brain Disorder Analysis by Using Deep Learning Based on DNA Sequence. Computer Systems Science and Engineering 2023;45(3):2447–2460.

[145] Anshori M, Mahmudy WF, Supianto AA. Classification Tuberculosis DNA using LDA-SVM. Journal of Information Technology and Computer Science 2019 Dec;4(3):233–240. http://dx.doi.org/10.25126/jitecs.201943113.

[146] Ali SN, Ahmed MT, Paul J, Jahan T, Sani SMS, Noor N, et al. Monkeypox Skin Lesion Detection Using Deep Learning Models: A Feasibility Study. ArXiv 2022;abs/2207.03342.

[147] https://doi.org/10.48550/arXiv.2207.03342.

[148] Chen J, Ma J, Han X, Zhou Y, Xie B, Huang F, et al. Rapid detection of hepatitis B virus DNA level based on interval-point data fusion of infrared spectra. Journal of Biophotonics 2023;16(3):e202200251. https://doi.org/10.1002/jbio.202200251.

[149] [143] Al Kindhi B. Optimization of Machine Learning Algorithms for Predicting Infected COVID-19 in Isolated DNA. International Journal of Intelligent Engineering & Systems 2020;13(4).

[150] [144] Mallick PK, Mohapatra SK, Chae GS, Mohanty MN. Convergent learning–based model for leukemia classification from gene expression. Personal and Ubiquitous Computing 2023;27(3):1103–1110. https://doi.org/10.1007/s00779-020-01467-3.