EDRAAK
Vol. **(2024)**, 2024, **pp**. 118–127
ISSN: 3078-8412

Review Article

Machine and Deep Learning Techniques in Cancer Prediction and Risk Stratification Using Bioinformatics in Big Data Era: A Review

Rana Abdulrahman Lateef[1, *]

[1] *Department of Cybersecurity Sciences, Baghdad College of Economic Sciences University, Baghdad, Iraq*

ARTICLE INFO

Article History
Received 20 Jun 2024
Revised: 12 Aug 2024
Accepted 12 Sep 2024
Published 3 Oct 2024

Keywords
Bioinformatics,
Machine Learning,
Deep Learning,
Cancer,
Big data.

ABSTRACT

Cancer remains a global health challenge, and early edge detection is crucial for improving patient outcomes and survival rates. Traditional diagnostic methods often face limitations in sensitivity and specificity, emphasizing the need to innovate approaches to enhance cancer diagnosis. This paper summarizes the recent applications and methodologies of Machine learning(ML) and/or Deep learning(DL) in bioinformatic for cancer prediction and diagnosis and how can be successfully employed to tackle problems such as patient classification, gene clustering, and biomarkers identification. This review constraints on three types of cancer: prostate, gastric, and colorectal.

1. INTRODUCTION

The cancer develops due to the improper regulation of cell growth and division. The human body consists of trillions of cells, which are the basic building blocks of life. These cells continually grow, divide, and undergo programmed cell death (apoptosis) to ensure the proper functioning of tissues. However, when this delicate balance is disturbed, it can result in the unrestrained increase of cells, leading to the genesis of cancerous tumors. Sometimes, there may be an abnormal cell growth, and that abnormal growth may lead to cancer. Nowadays, cancer poses a significant health burden in contemporary Australia. Projections indicate that approximately half of the Australian population will receive a cancer diagnosis by the age of 85. In 2021, the mortality of people increase of about 26,000 more than in 1981. This is basically due to the growth in population and aging. Even though a significant down of over 24.5% in the cancer mortality rate (deaths per 100,000 individuals), Australia still faces a substantial cancer burden. In 2024, an estimated 169,500 new cancer cases were diagnosed, and approximately 52,700 individuals succumbed to the disease[1]. Bioinformatics involves the integration of biology, computer science and information technology to analyze and make sense of data as in Figure (1), with applications spanning different fields. It is a rapidly growing and promising field which has emerged due to advances in biotechnology and data analysis techniques. it is considering as remarkable part of the informatics of traditional health, merging biomedical information with computer science. The comprehensive efforts and tasks carried out in bioinformatics highlight its critical role in analyzing and extracting valuable insights from the ever-expanding pool of biomedical data from various perspectives[2]. The basic and translational recent cancer research usually generates huge amount of data which is becoming increasingly reliant on calculating for their elucidation. These data obtained from various sources, such as epigenomics, next-generation sequencing of tumor DNA and RNA, imaging technologies, and histopathological evaluations. It is predictable that data-driven methods will soon be vital in clinical oncology, aiding in premature prognostic, more precise diagnoses, and improved disease administration. In response to this changing landscape, this volume offers a thorough overview of methods and tools for analyzing and interpreting cancer-related data, showcasing the latest advancements in cancer informatics. The book is set for a wide audience, including effective scientists in computational biology and bioinformatics, bioinformatics inventor, research and clinical oncologists looking for bioinformatics assistance, and cancer drug creator aiming to regulate their search for new composites [3].

*Corresponding author email: taught.rana.abdalrahman@baghdadcollege.edu.iq
DOI: https://doi.org/10.70470/EDRAAK/2024/015

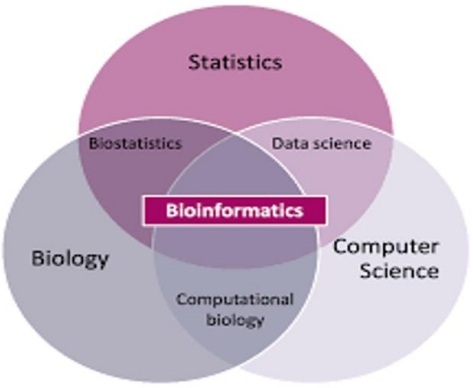

Fig. 1. The bioinformatics[2]

Bioinformatics, an interdisciplinary field and the core of recent monocular biology where computational techniques improved and used to alter biological data into knowledge and extract meaningful insights from it and translate them for biomedical applications. Machine learning, a powerful tool within artificial intelligence, has revolutionized bioinformatics by enabling the development of sophisticated models capable of analyzing vast datasets and uncovering intricate patterns[4]. The role of bioinformatics in cancer diagnosis such as Lung, Breast, Liver, Oral, Brain, and Ovarian have been discussed in many review research papers [5][6]. The rapid growth of algorithms, computing power, and the volume and velocity of data has transformed many industries and areas of research. This can help prepare businesses for these changes by using the collected data to make those changes and better adapt to meet their new consumer demand. Improved and fast-tracked data synthesis can massively help as industries from automotive, finance and healthcare to manufacturing[7]. Recently, Deep learning has become among the most vital and successful machine learning Techniques. this has been refreshed the cutting-edge performance of plenty of machine learning tasks and also make possible the development of various disciplines [8][39]. This paper aims at the survey of modern methods of ML and DL, explored in bioinformatics of cancer prediction and risk stratification. This review intends to summarize current progress in bioinformatics for carcinoma investigation. We present a preview of the ML techniques adopted, major results and prognostic accuracy of recent articles in the cancer detection settings. This review aims to provide useful insights and to act as a reference for future studies. Section 2 focuses on the applications of machine learning and bioinformatics to prognosis of cancers and Section 3 covers the place of deep learning and bioinformatics in this process. Finally, Section 4 demonstrate the study conclusions.

2. BIOINFORMATICS AND MACHINE LEARNING IN CANCER

The advancements in DL as a modern ML technology that can learn from complicate data without prior assumptions has been successfully applied in various bioinformatics studies, including drug-target interactions and drug synergy predictions[11]. DL is considered as subset of ML, has appeared as a vital technique motivated by the construction and function of the human brain. it is enabling the process of complex nonlinear relationship within data using artificial neural networks (ANNs) with multi-layers of interconnected nodes. This hierarchical form permits deep learning models to learn complicate patterns and features, conducted to state-of-the-art performance in numerous domains, including bioinformatics. speech, image recognition, natural language processing, and biomedical research are among the fields that DL has been successfully applied. furthermore, DL is considered as a powerful tool in these area since Its ability to extract high-level features from raw data, leverage distributed and parallel computing, and learn complex patterns without extensive manual intervention [12]. Several published articles tried to identify and solve different problems related to cancer using bioinformatics with machine Learning or with deep learning or as a hybrid.

Prostate Cancer (PCa)

Prostate Cancer Metastasis is the essential cause of mortality among patients. discovering narrative and powerful biomarkers is primary for realizing the mechanism of metastasis in PCa patients and promoting successful involvement[13].

He et al.[14] intended to seek for key genes and biological paths attached to Prostate cancer through utilizing bioinformatics technique by Differentially Expressed Genes (DEGs) which extracted from a dataset named GSE103512 and undergone Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) paths evaluation. In same regard, Hamzeh et al. [15] utilized a narrative ML technique to examine prostate tumors gene expression with various Gleason outcomes and determine possible genetic biomarkers for every Gleason set. The overtly accessible RNA-Seq dataset was acquired and classified patients depend on their Gleason outcome to produce a hierarchal representation for the progression of the disease. Cario et al. [16] demonstrated a narrative design strategy of ML guided dashboard for enhancing the tumor variants recognition in cfDNA. They first initiated a paradigm to categorize and assessment of candidate variants for inclusive on a targeted sequential panel, then this panel was employed to test tumor alteration in PC patients' cases with centralized disease in both in silico and hybrid capture settings. Wang et al. [17] identified possible multi-comics biomarkers for the premature recognition of the prediction recurrence of PC patients. An entire of 494 prostate adenocarcinoma (PRAD) patients (60-recurrent involved) from the Cancer Genome Atlas (TCGA) portal were examined by similar network fusibility and the auto-

encoder paradigm. IQBAL et al. [18] stated that images gained from a carcinoma patients are composed of necessary and complex attributes that is unable to be extracted effortlessly by classical diagnostic methods. Therefore, they utilized deep learning techniques which are fine-tuned and separate from hand-crafted attributes. The outcome was compared with hand-crafted traits like gray level co-occurrence matrix (GLCM), texture, and morphology by applying a non-deep learning classifier. Rammal et al. [19] studied the Gleason effect on several glands outcome from a novel optical microscopy method named SLIM. This novel optical microscopy mechanism merges the two traditional concepts in light imaging: Zernike's phase contrast microscopy and Gabor's holography. A machine learning technique were suggested to categorize these images into the corresponding Gleason outcome.ML methods stratified to homological persistence features effectively identify the correct Gleason outcome. Ying et al. [20] explored the genes with diagnostic amount in sick person with benign prostatic hyperplasia (BPH), discover the correlation among the immune microenvironment and the expression of diagnosis-related genes, and issue a molecular diagnosis reference and BPH immunotherapy. The differential expression of autophagy-related genes among BPH of a sick person and healthy controls was gained by differential examine. Next the genes associated to the diagnosis of BPH were scanned by a ML method and confirmed. Eventually, five vital genes (IGF1, PSIP1, SLC1A3, SLC2A1 and T1A1) were gained by Random Forest (RF) algorithm. Dai et al. [21] combined ML methods to promote a narrative mitophagy-related long non-coding RNA (lncRNA) signature for forecasting the advancement of PC. In utilizing the TCGA-PRAD data, they are identified a group of 4 key lncRNAs and formulate a risks core, exploring its possibility as a prognostic indicator. Table I summarizes the work related to prostate cancer diagnosis.

TABLE I. SUMMARIZATION OF WORKS RELATED TO PROSTATE CANCER DIAGNOSIS

Author	Data	Method	Key finding(s)
He et al. 2019 [14]	PCa microarray expression dataset GSE103512	bioinformatic evaluation involving GO, (KEGG) enrichment, PPI network, hub gene identification, and module analysis	Entirely, 252 (186 upregulated - whereas 66 downregulated) of Differentially expressed genes (DEGs) were identified KLK3, CDHI, FOXA1, and EPCAM - could possibly work as powerful and definitive molecular biomarkers for PCa. prediction
Hamzeh et al 2019 [15]	-RNA-Seq data of a cohort consist of 104 PC sick person from (NCBI). - (GEO)	-A novel machine learning method was applied for analysis-Seq dataset from NCBI's GEO repository was utilized. A hierarchical model with standard classifiers was developed. - Class disparity and hybrid feature selection methods were employed. -Naive Bayes and SVM classifiers were used for predictions. Synthetic minority oversampling method (SMOTE) was employed to solve imbalance in class.	A model achieved 93.3% accuracy with first dataset. Validation on the second dataset yielded 87% accuracy. Six gene transcripts were identified as differentially expressed. Naive Bayes classifier outperformed others in accuracy. Gleason score 6 identified with 100% accuracy. PIAS3 and UBE2V2 identified as potential biomarkers.
Cario et al. 2020 [16]	-ICGC dataset: derived from the International Cancer Genome Consortium (ICGC) which included Whole Genome Sequence (WGS) tumor variant data from 550 PC patients. - UCSF Cohort dataset: utilized data from 23 patients derived from California University, San Francisco (UCSF).	-Machine learning model for variant classification and scoring. -In silico screening of tumor variants from PC patients. -Hybrid acquisition and sequencing of cfDNA at 2500X depth. -Development of a targeted sequencing panel for mutations.	-A novel machine-learning panel improved tumor variant detection. -The panel outperformed two existing designs in in silico tests. -Detected tumor diverse in total of 18 prostate cancer patients. -Identified mutations in known driver genes like HRAS. -Machine learning optimized targeted sequencing panel composition. -The approach enhances sensitivity for early cancer detection.
Wang et al. 2021 [17]	A whole of 494 prostate adenocarcinoma (PRAD) patients (involved of 60 repetitive) from the Cancer Genome Atlas (TCGA)	-Autoencoder model for feature reduction and analysis. -Similarity Network Fusion (SNF) for clustering and prognostic. -Univariate Cox regression examine for recurrence association. -K-means clustering for sample grouping. -Spectral clustering for SNF sample clustering.	-Six omics biomarkers were identified: TELO2, ZMYND19, miR-143, miR-378a, cg00687383, cg02318866. -Multiomics panel achieved p-value = 2.97×10^{-15}. -Five-year recurrence prediction performance had AUC = 0.789. -utilizing autoencoder with SVM classifier results in 97.1% of accuracy.

			-Two distinctive repetition -risk subgroups were recognized in TCGA data.
IQBAL et al. 2021 [18]	the dataset consists of 230 for MRI scans of sick person with variant classes and representation. The dataset was scored by the Health Insurance Portability and Liability Act 1996 (HIPPA) regulations.	- Deep learning methods: Residual Net (ResNet − 101) and Long Short-Term Memory (LSTM) -ML techniques: SVM, Gaussian Kernel, k-nearest neighbor-Cosine (KNN − Cosine), Naive Bayes kernel, DT , and RUS Boost tree.	- using KNN-Cosine with accuracy (99.07%). -The LSTM method output performance with accuracy (99.48%), -exploit DL technique ResNet − 101, yield 100% Accuracy and AUC =1 for Kernel Naive Bayes, SVM, Gaussian and RUSBoost Tree. - the outcomes reveal that ResNet − 101 outperformed than non-DL methods and LSTM.
Rammal et al. 2022 [19]	500 images of prostate cancer biopsies, each with a size of 10,000 x 10,000 pixels. Each image was composed of glands that averaged around 1,000 x 1,000 pixels in size	-Persistent-Homology-Based Machine Learning (PHML) for feature extraction. -Supervised machine Learning: RF, MN, SVM, LDA, DT, and Naïve Bayes for classifying grand image. -Spatial Light Interference Microscopy (SLIM) for imaging	- DT classifier yield accuracy of 99.3%. - RF classifier yield accuracy of 80.1%. - SVM classifier yield accuracy of 84.5%. - Naïve Bayes classifier yield accuracy of 90.8%. - linear discriminant classifier yield accuracy 73.2%.
Ying et al. 2023 [20]	GEO database. **GSE6099 Dataset**: Contains 21 BPH samples and 4 control samples. **GSE119195 Dataset**: Comprises 5 BPH samples and 3 control samples	-Differential expression analysis using limma algorithm. - Machine learning algorithm for gene screening. - GO and KEGG enrichment analysis performed. - CIBERSORT algorithm for immune cell proportion calculation.	-identified 125 differentially expressed autophagy-related genes. -Obtained five key diagnostic genes using random forest algorithm. -Analyzed immune cell infiltration related to diagnostic genes. -Established PPI networks for significant autophagy-related genes. -Conducted GO and KEGG enrichment analyses on DEGs. -Found significant correlation between IGF1 and M2 macrophages.
Dai et al.2024 [21]	Cancer Genome Atlas (TCGA) database	-ML methods to promote a mitophagy-related lncRNA signature. -Statistical Analysis: Various statistical methods were utilized to analyze the data. - Student T-test was employed for continuous variables that spreads normally. - Mann-Whitney U-test was performed for continuous variables that non-ordinary spread	The study successfully developed a narrative mitophagy-related long non-coding RNA (lncRNA) signature that serves as a prognostic tool for forecasting the progression of PC.

2.1 Gastric Cancer (GC)

Gastric Cancer is among the majority widespread worldwide cancers. The molecular techniques of GC are still indistinct and not fully realized. GC situations are mainly detected at an advanced stage, revealing poor prediction. The development in molecular biology mechanisms permits us to gain deeper comprehension of accurate molecular techniques and permit us to recognize the key-genes in the progression and carcinogenesis of GC. Shon et al. [22] proposed a classification technique leveraging DL and determine its performance to the gene term data gained from GC sick-person. They merged the RNA-seq gene ex-pression data with clinical data, searched candidate genes, and examined them by applying Convolutional Neural Network (CNN) algorithm. During their work, they carry out learning by the type of sample and essential level of stomach cancer patients and confirm the outcomes. Gilani et al. [23] searched for determine possible miRNAs for Gastric cancer by using GSE106817 data with 2,566 miRNAs to train the ML paradigms. They utilized the Boruta ML variable election method to recognize the powerful miRNAs attached with GC in the training sample, then confirmed the prognoses models in the separated data sample of GSE113486. Eventually, an ontological examination was performed on recognized miRNAs to elicit the pertinent relation. Chen et al. [24]explored the possible hub genes of GC accompanied by a diagnostic estimate. The narrative biomarkers were exposed across many databases of gastric cancer–related genes. The hub genes (ESRRG, ATP4A, and ATP4B) were exposed by merged of weighted gene co-expression network analysis (WGCNA), gene–gene interaction network analysis, and supervised attributes selection methods. ML techniques involving data preprocessing, cross-validation, and model selection. achievement assessments were tested on the hub-gene expression profiles in five Gene

Expression Omnibus datasets and confirmed on a GEO external validation (EV) dataset. Niu et al. [25] utilized a data from the GEO dataset to show DEGs among normal gastric tissues and GC. KEGG and GO enrichments were applied to examine the task of DEGs and the STRING database and Cytoscape s/w to create protein–protein network and discover hub genes. The expression levels of hub genes were assessed by employed TCGA database. Talebi et al. [26] developed predictive models by employing different ML classifiers depend on both clinical variables and demographic to forecast metastasis status of patients with Gastric Cancer. Azari et al. [27] identified possible diagnostic and prediction of miRNAs in Gastric Cancer with the implementation of ML methods. To detect popular molecular techniques of the miRNAs, the targets of joint gene were detected by online databases like miRWalk, Targetscan, and miRDB, Functional and fertility analyzes were applied through utilizing Kyoto Database of Genes and Genomes (KEGG) and Gene Ontology (GO), also recognition of protein–protein interactions (PPI) through STRING database. Xie et al. [28] employed ML to determine GC diagnostic genes and explore their relation with immune cell infiltration. Suriya et al. [29] tested the Transcriptome profiles of Gastric Cancer patients to determine DEG among the tumor and adjacent normal tissues. later, they built networks for protein–protein interaction to discover the important hub genes. Besides the bioinformatics incorporation of ML Techniques like SVM, the recursive attributes elimination was utilized to choose the maximum informational genes. Table II summarizes the work related to Gastric cancer diagnosis.

TABLE II. SUMMARIZATION OF WORKS RELATED TO GASTRIC CANCER DIAGNOSIS

Author	Data	Method	Key finding(s)
Shon et al. 2019 [22]	TCGA, a worldwide cancer database contains data of 60,483 genes collect from 334 sick person having Gastric Cancer.	-Principal Component Analysis (PCA) for extracting features. - (CNN) for classification. -Heatmap generation for data visualization.	-classified sample type with 95.96% accuracy. -Obtained 50.51% accuracy for important status classification. -Identified 320 principal components from gene expression data. -Generated heatmap images for CNN input. -The method can predict stomach cancer prognosis effectively. -Increased data can resolve overfitting issues. -High accuracy achieved for sample type classification.
Gilani et al. 2022 [23]	GSE106817 dataset consists of 2,566 miRNAs data gained from 2,759 noncancer controls, and 115 GC situations (4%).	- Machine learning variable selection method named Boruta was utilized. -Five machine learning algorithms: RF, LR, DT, XGBT, ANN. -Synthetic Minority Oversampling method (SMOTE) was implemented for regulate the training data. -GeneCodis tool was used for ontological analysis of miRNAs. -Decision trees were employed for classification with a cut-off point.	-115 out of 2,874 patients had gastric cancer. -30 miRNAs identified as potential biomarkers for gastric cancer. -hsa-miR-1343-3p ranked maximum among identified miRNAs. -hsa-miR-1343-3p predicted gastric cancer with 100% precision. -Ontological analysis confirmed strong relationships with cancer-associated genes.
Chen et al. 2022 [24]	-6 datasets from the Gene Expression Omnibus (GEO). include GSE19826, GSE27342, GSE29272, GSE54129, GSE66229. -GSE66229 was used for weighted gene coexpression network analysis. -GSE33335 served as an independent dataset for external model assessment.	-Bioinformatics methods for gene selection and analysis. -Weighted gene co-expression network analysis (WGCNA). - analysis of Gene-gene interaction network. -Supervised attributes selection methods. -Machine learning (ML) techniques for diagnostic models. -Stratified k-fold cross-validation. -Random permutation validation. -Immunohistochemistry for gene expression verification.	-3 hub genes identified: ATP4A , ESRRG, and ATP4B. -Support vector machine model showed highest diagnostic performance. -Achieved 0.93 AUC on test dataset, 0.99 on validation. -Semi-supervised model also demonstrated strong predictive ability. - outcomes indicate mild impact on auxiliary diagnosis. -Comprehensive analysis enhances objectivity of diagnostic model. -Hub genes validated through multiple bioinformatics methods.
Niu et al. 2022 [25]	The dataset of RNA-sequencing consisting of GC tissue swabs as well as ordinary tissue swabs gained from the GEO dataset and three GEO datasets, including GSE19826GSE54129, and GSE118916.	-Differentially expressed genes (DEGs) were screened among GC as well ordinary tissues. -GO and KEGG fertility analyzed the function of DEGs.	-Identified 607 differentially expressed genes (DEGs) in gastric cancer. -DEGs mainly enriched in extracellular matrix and integrin binding.

		-STRING dataset and Cytoscape s/w generated protein-protein networks	-Four hub genes proposed: COL1A1, COL5A2, P4HA3, SPARC. -COL1A1, COL5A2, P4HA3, SPARC show high diagnostic and prognostic value. -Survival analysis linked hub genes to overall survival of GC. -ROC evaluation verified max value for diagnostic of SPARC. -Correlation analysis showed strong interrelation among the four hub genes. - recognized about 294 up-regulated whereas 313 down-regulated genes in GC.
Talebi et al. 2023 [26]	-The study used a dataset of 733 gastric cancer patients. - The dataset included 10 features related to demographics and clinical characteristics. - Data was collected from electronic records at Taleghani tertiary hospital.	-Machine learning classifiers: - Naive Bayes - (SVM) -Neural Network (NN) -Decision Tree (DT) -Logistic Regresion (LR) - Random Fores (RF)	-SVM is the top-performing machine learning model. -NN and RF are also effective algorithms. -Tumor size and age are crucial variables in RF model.
Azari et al.2023 [27]	-TCGA database - miRWalk, miRDB, and Targetscan.databases - Kyoto of Genes and Genomes (KEGG) and Gene Ontology (GO) database. -STRING database	-ML techniques: SVM, Random Forest, k-NN, Logistic Regression, DTS. -Data analysis using TCGA database. Heatmap analysis for feature selection. -ROC curve for diagnostic evaluation. -evaluation of Protein-protein interaction network by utilizing STRING dataset.	Among the ML algorithms, SVM was chosen (AUC:88.5%, Accuracy:93% in GC). -A panel of 29 miRNAs was identified as potential biomarkers. -the hsa-miR-21, hsa-miR-133a, hsa-miR-146b, and hsa-miR-29c showed high prediction power.
Xie et al 2023 [28]	**-GEO Datasets**: GSE13911, GSE15459, GSE19826, GSE54129 GSE79973 These five datasets were merged to form the training set, which included a total of 371 samples for GC. Furthermore, 77 samples of normal gastric tissue, after removing batch effects.	-Machine learning algorithms: LASSO and SVM-RFE. -Bioinformatics analysis of gastric cancer datasets. -Receiver operating characteristic (ROC) curve evaluation. -CIBERSORT for immune cell infiltration analysis.	-Eight candidate diagnostic genes for gastric cancer were identified. -Six diagnostic genes showed significant prognostic value. -Diagnostic genes are more correlated to immune cell infiltration. -Bioinformatics & machine learning methods were effectively utilized.
Suriya et al. 2023 [29]	-PRJNA555737 and PRJNA435914 acquired from the National Center for Biotechnology Information-Sequence Read Archive database - (PRJNA555737) compose of 12 swabs (six tumors and six adjacent ordinary tissues) & PRJNA435914 compose of sixty-eight swabs. - profiling transcriptome of Stomach Adenocarcinoma (STAD) and their conformable clinical data were summed from the TCGA database. it consist of 407 swabs (33 normal and 376 tumor tissues)	-Protein-protein interaction (PPI) network construction. The Network Analyzer (ver. 4.4.8) bundle is utilized to evaluate network for the estimation of the degree and betweenness centrality. - (SVM) with recursive feature elimination (RFE) technique was applied for attributes selection of diagnostic genetic markers from the important genes linked with Gastric Cancer.	-Identified 160 significant genes related to gastric cancer. -Found about88 upregulated and 72 downregulated genes. -Discovered 10 hub genes through protein-protein interaction networks. -KIF14 and TRIP13 are potential diagnostic biomarkers. -TRIP13 overexpression linked to poor prognosis in gastric cancer. -High diagnostic value for KIF14 and TRIP13 in ROC analysis. -KIF14 and TRIP13 are potential diagnostic biomarkers for gastric cancer. -The study aids understanding of gastric cancer pathophysiology. -Integrated bioinformatics approaches revealed significant gene candidates.

2.2 Colorectal cancer (CRC)

CRC is a 3rd majority widespread tumor all over the world. It is approximated that the universal encumbrance of CRC will raise to more than 2.2 million new instances and 1.1 million yearly deceases by 2030 [30]. LI et al. [31] used the data of gene mutation, gene expression and gene relation to build attribute vectors of gene pattern and exploit it to learn classification models for pattern identification and to mine seven colorectal cancer-related potential driver genes. The aim is not only to confirm the advantage of extracting characteristics but considered it as a significant point to estimate the growth and alteration of colorectal cancer, and to demonstrate that employing gene network techniques to mine driver genes is remarkably more suitable than mining techniques without utilizing gene networks. Hammad et al. [32] studied the diagnosis of CRC by integrating bioinformatics and machine learning to recognize and validate potential biomarkers from gene expression data. Gene Ontology (GO) and Kyoto Enrichment of Genes and Genomes (KEGG) evaluation were applied to recognize biological procedures associated with the DEGs. Lacalamita et al. [33] focused on CRC carcinogenesis by develops a prognostic-classifier for adenoma-carcinoma sequence by microarray gene expression profiles of essential adenoma, CRC, and ordinary colon epithelial tissues. expression profiles of four genes from the Gene Expression Omnibus database, consisting of 465 swabs were preprocessed to recognize (DEGs) among adenoma tissue and essential CRC. The classification was performed with machine learning algorithms. Acharjee et al. [34] evolved a narrative ML to survey CRC gene corporation. various ML techniques were applied as classifiers to recognize genes that can be utilized as diagnostics for CRC using gene expression and clinical data. Gene ontology enrichment evaluates these differentially expressed genes (DEGs) were applied and prognosticated that gene signatures were attached with miRNAs. Xue et al. [35] identifies key genes linked to ferroptosis in CRC. Machine learning and bioinformatics evaluation were utilized for gene identification. Eleven ferroptosis-related differentially expressed genes were identified. Four hub genes were determined: TFR2, NOX4, ALOXE3, and CA9. Bostanci et al. [36] explored the integration of omics technologies, particularly transcriptomics, with ML and DL to enhance diagnostic and prognostic capabilities for CRC. The research focused on using RNA-seq data from circulating extracellular vehicles (EVs) to predict colon cancer and classify its stages, leveraging the unique RNA profiles found in tumor-derived EVs and utilizing log2-transformed RPM values for reliable RNA-seq analysis, focusing on comparing healthy individuals with cancerous patients. miRNA isoform and exRNA stability analyses were conducted to enhance the understanding of RNA profiles relevant to colon cancer. Five canonical ML methods and three DL algorithms were used to pronounce the colon cancer and its stages. Vaziri-Moghadam et al. [37] identified diagnostic gene biomarkers for CRC by evaluating DEG from tumor and ordinary swabs. Nine candidate genes were recognized using LASSO logistic regression. Liang et al. [38] investigated E3 ubiquitin ligase-associated genes and colon cancer. Gene expression profiles and clinical data were analyzed. Two molecular clusters of E3-related genes were identified. A prediction model was built using machine learning techniques. Table III summarizes the works related to colorectal cancer diagnosis.

TABLE III. SUMMARIZATION OF WORKS RELATED TO COLORECTAL CANCER DIAGNOSIS

Author	Data	Method	Key finding
LI et al. 2020 [31]	CGC database NCBI (PubMed) database	-Machine learning methods: weighted KNN, Weighted Naïve Bayes, and Multi-level SVM for classification models. -Signed random walk restart method ranks nodes in signed networks. -Gene mutation and expression data are analyzed for feature extraction.	-The study successfully identifies potential colorectal cancer driver genes. -Gene network methods outperform non-network methods in predictions. -Structural features significantly enhance classification ability. -Feature fusion improves predictive ability for driver genes.
Hammad et al. 2021 [32]	-The gene expression microarray GSE103512 dataset from GEO comprised of a entirely 69 swabs (57 colorectal cancers with 12 ordinary swabs). -TCGA and GEPIA datasets. For validation	-Bioinformatics analysis of gene expression microarray data. -Machine learning techniques for biomarker identification. -Differentially expressed genes (DEGs) identification and functional analysis. -Protein-protein interaction (PPI) network analysis using STRING database. -Support Vector Machine (SVM) for diagnostic value prediction. -Receiver Operating Characteristic (ROC) curve analysis for biomarker evaluation. -Kaplan-Meier survival analysis for prognostic assessment.	-Identified 105 differentially expressed genes (DEGs) for CRC. -Ten hub genes were determined as potential biomarkers for CRC. -Four hub genes correlated with CRC tumor stages. -ROC curve AUC exceeded 0.92 for biomarker prediction. -Survival analyses confirmed prognostic values of hub genes. -Functional enrichment showed DEGs involved in cancer progression processes.

		-ROC curve analysis for specificity and sensitivity evaluation.	
Lacalamita et al. 2021[33]	The raw microarray data was sourced from the Gene Expression Omnibus (GEO) database, which is a public repository for gene expression data. The merged dataset consists of a entirely 465 samples, classified into three seperate groups: **-Healthy Controls**: 105 samples from individuals without any colorectal abnormalities. **-Adenoma Group**: 155 samples from patients with adenomatous polyps, which are precursors to cancer. **-CRC Group**: 205 samples from patients diagnosed with colorectal cancer.	-Microarray gene expression profiling. -Differential expression analysis using unpaired t-test. -Feature selection via Boruta algorithm and Stepwise Regression. -K-Means clustering method. -Machine learning algorithms: LM, RF, k-NN, ANN.	-11,530 genes identified as differentially expressed. -240 important genes selected using the Boruta algorithm. -56 highly important genes determined for classification. -k-NN model achieved 91.11% accuracy on validation cohort. -Six DEGs identified related to patient prognosis. -Potential biomarker for early CRC diagnosis proposed.
Acharjee et al. 2022 [34]	GEO Dataset: GSE44861 GSE20916 GSE113513	-LR, naive Bayes. - Adaboost ,ExtraTrees . -Random Forest and XGBoost. -Enrichment analysis was performed on gene signatures. -Fivefold cross-validation, bootstrapping, and LOOCV were used to prevent overfitting.	- LR with an accuracy of 96.4% using GSE44861 as training data and GSE113513 as testing data. -The Random Forest with an accuracy of 98.2%. -The Extra Tree classifier showed preferable achievement when GSE113513 was used as training data and GSE44861 as testing data. -Naïve Bayes Classifier When using GSE20916 as training data and GSE44861 as testing data yielded an accuracy of 90.1%. - Logistic Regression achieved using GSE20916 as training data and GSE113513 as testing data, better performance by yielding an AUROC of **99%**. - When GSE44861 was used as training data and GSE113513 as testing data. The study successfully identified novel gene associations with CRC that could serve as diagnostic markers, emphasizing the potential of machine learning in translational research for cancer diagnostics
Xue et al. 2023 [35]	(GEO) datasets for CRC from the National Center for Biotechnology Information (NCBI).	-LASSO regression and SVM models were built. -Immune infiltrates were identified using the CIBERSORT algorithm. -Correlation analyses were performed using Spearman and Pearson algorithms. -Machine learning techniques were applied for gene identification.	Recognized 11 ferroptosis-related differentially expressed genes (DEGs) in CRC. -Four hub genes: NOX4, TFR2, ALOXE3, CA9. -NOX4 expression correlates with immune cell infiltration. -Low NOX4 levels linked to favorable patient prognosis. -Machine learning models demonstrated excellent diagnostic ability.
Bostanci et al. 2023 [36]	The dataset is a well-structured collection of RNA-seq data includes a total of 300 samples, which are divided into healthy and cancerous categories. Initially, 50 healthy samples were randomly selected, and to augment this, 50 additional healthy samples were generated, resulting in 100 healthy samples in total. For the cancerous samples, 100 augmented cancerous	-Canonical ML classifiers: kNN, LMT, RT, RC, RF. -DL paradigm: 1-D CNN, LSTM, BiLSTM. -Genetic algorithm for hyper-parameter optimization. -Min-max normalization for data preprocessing. -Feature selection to identify informative exRNA transcripts.	-Canonical ML techniques achieved 97.33% accuracy in predictions. -1-D CNN model reached 97.67% accuracy in cancer prediction. -BiLSTM model achieved 98% accuracy in cancer stage classification. -RC, LMT, and RF performed best in cancer prediction. -Feature selection improved model accuracy and reduced training time.

	pattern were generated from a randomly selected cancerous sample, leading to a whole of 200 cancerous patterns.		-McNemar's test indicated statistically significant performance differences among models.
Vaziri-Moghadam et al. 2024 [37]	The GEO database	-Differential gene expression analysis utilizing 'limma' bundle. -Gene co-expression network analysis with 'CEMiTool' package. -LASSO logistic regression for candidate gene screening. -ML : SVM, RF, GBM, method and ANN.	-Identified 283 differentially expressed genes (DEGs) in CRC. -Eleven candidate diagnostic genes were recognized using LASSO logistic regression. -Nine genes showed AUROC values over 0.92 in validation sets. -All machine learning algorithms achieved AUROC scores above 0.95.
Liang et al. 2024 [38]	TCGA, GTEx, GSE17537 and GSE29621 databases	-Gene expression profiles and clinical data were acquired. -Coexpression analysis identified E3-related genes (ERGs). -Weighted gene coexpression network analysis (WGCNA) was conducted. -Differential expression analysis was performed. -Consensus clustering recognized two molecular clusters. -Cox regression evaluation was conducted. -Prognostic model constructed using 10 machine learning algorithms.	-Two E3-related gene clusters identified in colon cancer. -Cluster A shows better prognosis than cluster B. -Prognostic model validated in internal and external datasets. -Significant immune infiltration differences between risk groups observed. -High-risk group has lower IC50 for some antitumor drugs. -Ectopic PRELP expression inhibits CRC cell migration and proliferation.

3. CONCLUSIONS

This study shows a comprehensive overview of the current works on cancer prediction using ML and DL approaches with bioinformatics. The implementation of these techniques in bioinformatics has escort in a new era for cancer prediction and diagnosis, offering unique potential for improving early detection, diagnostic accuracy, and personalized treatment strategies. This review underscores the transformative impact of ML and DL algorithms in analyzing complex genomic, proteomic, and clinical data, enabling the identification of subtle patterns and biomarkers associated with cancer. These advancements facilitate the development of predictive models with high precision, surpassing traditional methods in their ability to handle large-scale and heterogeneous datasets. Despite these promising developments, challenges persist. Issues such as data scarcity, imbalanced datasets, and model interpretability continue to hinder the translation of these technologies into clinical practice. recording these limitations needs to leverage of robust, explainable algorithms and the integration of diverse, high-quality datasets. Furthermore, interdisciplinary collaboration among clinicians, bioinformaticians, and data scientists is essential to ensure the clinical relevance and usability of these models. Ethical considerations, including data privacy and algorithmic fairness, must also be prioritized to foster trust and widespread adoption. In conclusion, ML and DL hold immense promise for revolutionizing cancer prediction and diagnosis. By addressing existing challenges and fostering innovation, these technologies can remarkably improve the precision and efficiency of cancer care, paving the way for a future where premature detection and personalized treatment are accessible to all.

Funding:

No external financial assistance or institutional funding was utilized for conducting this research. The authors assert that all research-related activities were self-financed.

Conflicts of Interest:

The authors declare that there are no competing interests associated with this work.

Acknowledgment:

The authors would like to thank their institutions for their steadfast encouragement and logistical support throughout this research journey.

References

[1] D. M. Parkin, F. Bray, J. Ferlay, and P. Pisani, “Global cancer statistics, 2002,” CA: A Cancer Journal for Clinicians, vol. 55, no. 2, pp. 74–108, 2005. DOI: 10.3322/canjclin.55.2.74.
[2] P. R. Said, “Open access development of new diagnostic methods for early detection of cancer: Integration of genomic and bioinformatics technologies,” pp. 0–6, 2023.
[3] A. Krasnitz, *Cancer Bioinformatics*. 1878.
[4] A. Lidia, Q. Cavalcante, M. D. B. Braga, and R. B. Kato, *Advances in Bioinformatics*, no. August, 2021.

[5] V. M. Nelakurthi, P. Paul, and A. Reche, "Bioinformatics in early cancer detection," *Cureus*, vol. 15, no. 10, pp. 4–9, 2023, doi: 10.7759/cureus.46931.
[6] J. H. Valand, M. Kyomukamaa, R. Atino, and A. G. Nabwami, "Role of bioinformatics in cancer diagnosis," pp. 1–16.
[7] F. Bajaber et al., "Big data 2.0 processing systems: Taxonomy and open challenges," *J. Grid Comput.*, vol. 14, pp. 379–405, 2016.
[8] Y. Li, "Towards structured prediction in bioinformatics with deep learning," 2020.
[9] S. Wan, Y. Fan, C. Jiang, and S. Li, *Bioinformatics and Machine Learning for Cancer Biology*. MDPI, 2022.
[10] D. Painuli and S. Bhardwaj, "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review," *Comput. Biol. Med.*, vol. 146, p. 105580, 2022.
[11] H. Li et al., "Modern deep learning in bioinformatics," *J. Mol. Cell Biol.*, vol. 12, no. 11, pp. 823–827, 2020.
[12] K. Lan et al., "A survey of data mining and deep learning in bioinformatics," *J. Med. Syst.*, vol. 42, no. 8, p. 139, Jun. 2018, doi: 10.1007/s10916-018-1003-9.
[13] B. Alizadeh, H. Asadzadeh, and A. Behmanesh, "A machine learning approach identified a diagnostic model for pancreatic cancer through using circulating microRNA signatures," *Pancreatology*, vol. 20, no. 6, pp. 1195–1204, 2020, doi: 10.1016/j.pan.2020.07.399.
[14] Z. He, X. Duan, and G. Zeng, "Identification of potential biomarkers and pivotal biological pathways for prostate cancer using bioinformatics analysis methods," pp. 1–21, 2019, doi: 10.7717/peerj.7872.
[15] O. Hamzeh et al., "A hierarchical machine learning model to discover Gleason grade-specific biomarkers in prostate cancer," *Diagnostics (Basel, Switzerland)*, vol. 9, no. 4, Dec. 2019, doi: 10.3390/diagnostics9040219.
[16] C. L. Cario et al., "A machine learning approach to optimizing cell-free DNA sequencing panels: with an application to prostate cancer," pp. 1–9, 2020.
[17] T. Wang et al., "Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion," 2021.
[18] S. Iqbal et al., "Prostate cancer detection using deep learning and traditional techniques," *IEEE Access*, vol. 9, pp. 27085–27100, 2021, doi: 10.1109/ACCESS.2021.3057654.
[19] A. Rammal et al., "Machine learning techniques on homological persistence features for prostate cancer diagnosis," *BMC Bioinformatics*, pp. 1–22, 2022, doi: 10.1186/s12859-022-04992-5.
[20] A. Ying, Y. Zhao, and X. Hu, "Identification of biomarkers related to prostatic hyperplasia based on bioinformatics and machine learning," *Math. Biosci. Eng.*, vol. 20, no. March, pp. 12024–12038, 2023, doi: 10.3934/mbe.2023534.
[21] C. Dai et al., "Machine learning-based integration develops a mitophagy-related lncRNA signature for predicting the progression of prostate cancer: a bioinformatic analysis," *Discov. Oncol.*, 2024, doi: 10.1007/s12672-024-01189-5.
[22] H. S. Shon et al., "Classification of stomach cancer gene expression data using CNN algorithm of deep learning," vol. 20, no. 1, pp. 15–20, 2019.
[23] N. Gilani et al., "Identifying potential miRNA biomarkers for gastric cancer diagnosis using machine learning variable selection approach," *Front. Genet.*, vol. 12, no. January, pp. 1–10, 2022, doi: 10.3389/fgene.2021.779455.
[24] Q. Chen et al., "ESRRG, ATP4A, and ATP4B as diagnostic biomarkers for gastric cancer: A bioinformatic analysis based on machine learning," *Front. Physiol.*, vol. 13, no. June, pp. 1–14, 2022, doi: 10.3389/fphys.2022.905523.
[25] X. Niu et al., "Identification of potential diagnostic and prognostic biomarkers for gastric cancer based on bioinformatic analysis," *Front. Genet.*, vol. 13, no. March, pp. 1–12, 2022, doi: 10.3389/fgene.2022.862105.
[26] A. Talebi et al., "Predicting metastasis in gastric cancer patients: machine learning-based approaches," *Sci. Rep.*, pp. 1–12, 2023, doi: 10.1038/s41598-023-31272-w.
[27] H. Azari et al., "Machine learning algorithms reveal potential miRNAs biomarkers in gastric cancer," *Sci. Rep.*, pp. 1–12, 2023, doi: 10.1038/s41598-023-32332-x.
[28] R. Xie et al., "Identification of the diagnostic genes and immune cell infiltration characteristics of gastric cancer using bioinformatics analysis and machine learning," *Front. Genet.*, no. January, pp. 1–15, 2023, doi: 10.3389/fgene.2022.1067524.
[29] V. Suriya et al., "Gastric cancer biomarker candidates identified by machine learning and integrative bioinformatics," *OMICS*, vol. 27, no. 6, pp. 260–272, 2023, doi: 10.1089/omi.2023.0015.
[30] M. Ahmed, "Colon cancer: A clinician's perspective in 2019," vol. 13, no. 1, pp. 1–10, 2020.
[31] Y. Li et al., "Screening of pathogenic genes for colorectal cancer and deep learning in the diagnosis of colorectal cancer," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3003999.
[32] A. Hammad, M. Elshaer, and X. Tang, "Identification of potential biomarkers with colorectal cancer based on bioinformatics analysis and machine learning," *Math. Biosci. Eng.*, vol. 18, no. August, pp. 8997–9015, 2021, doi: 10.3934/mbe.2021443.
[33] A. Lacalamita et al., "A gene-based machine learning classifier associated with the colorectal adenoma–carcinoma sequence," pp. 1–14, 2021.
[34] A. Acharjee, "Machine learning-based identification of colon cancer candidate diagnostic genes," pp. 1–15, 2022.
[35] F. Xue, J. Jiang, and J. Kou, "Screening of key genes related to ferroptosis and a molecular interaction network analysis in colorectal cancer using machine learning and bioinformatics," *J. Gastrointest. Oncol.*, vol. 14, no. 3, pp. 1346–1359, 2023, doi: 10.21037/jgo-23-405.
[36] E. Bostanci et al., "Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer," 2023.
[37] S. Zhang *et al.*, "Machine learning-based identification of diagnostic biomarkers in colorectal cancer using gene expression data," *Front. Genet.*, vol. 13, pp. 1–12, 2022.
[38] L. Liang, X. Liang, X. Yu, and W. Xiang, "Bioinformatic analyses and integrated machine learning to predict prognosis and therapeutic response based on E3 ligase-related genes in colon cancer," *J. Cancer*, vol. 15, 2024, doi: 10.7150/jca.98723.
[39] A. Alsolami et al., "AI-driven cybersecurity solutions for healthcare data protection: Challenges and future directions," IEEE Access, vol. 10, pp. 1–14, 2022.