

A Review on Anisotropic Anomaly based Cancer-Causing Gene Detection using Hybrid ML Techniques

Thulasi Bai A¹ and T. Sam Pradeepraj²

¹⁻²Kalasalingam Academy of Research and Education, Anandnagar, Krishnankoil-626126, Tamil Nadu, India
Email: mail2thulasibai@gmail.com, t.sampradeepraj@klu.ac.in

Abstract—Cancer refers to a set of illnesses brought on by uncontrollable cell development that can affect various body organs. According to the World Health Organization (WHO), cancer is the second leading cause of death after cardiovascular illnesses. Gene expression can play a critical role in the early detection of cancer since it can provide information about an organism's genetic composition as well as the biochemical processes occurring in tissue and cells. This study presents a review on cancer causing gene detection based on dataset analysis, Hybrid Radiomics solutions-based DL model, semi-supervised ensemble learning, Anisotropic Anomaly-based Gene detection techniques, early diagnosis of cancer causing genes, and performance analysis of various cancer causing gene techniques. This review provides a clear idea of strategies used in cancer causing gene detection. Additionally, it offers a comparison study to show the efficacy of different methods in cancer-causing gene discovery methodologies. Moreover, future directions are provided to overcome the existing impact of various techniques in the detection of cancer causing genes.

Index Terms— carcinoma, machine learning, gene expression analysis, semi-supervised learning, deep learning, radiomics solution.

I. INTRODUCTION

Cancer is a fatal condition that is typically brought on by the aggregation of numerous pathological changes and inherited illnesses [1]. The human body can develop cancerous cells everywhere, which are abnormal, life-threatening growths. To ascertain what might be useful for its therapy, the early stages of cancer, also referred to as tumors, must be rapidly and precisely diagnosed. Even though every modality has its special considerations, some of the main causes of mortality include challenging histories, inadequate diagnoses, and subpar treatment. to gather and discuss the latest advances in machine learning for the identification of leukaemia, as well as cancers of the breast, brain, lung, and other parts of the human body [2]. The knowledge of the links between infections and cancer has advanced significantly over the past 60 years, and this has been accompanied by an increase in the chance of developing cancer. Over the past 60 years, significant advances in understanding the links between infections and cancer have identified 11 group 1 carcinogens, with HBV, high-risk HPV, HCV, and Helicobacter pylori causing over 90% of infection-related cancers, prompting bold calls for high-level action from the scientific and public health communities [3-4]. Numerous cancers are directly correlated with alcohol use. Alcohol use contributes significantly to the global disease burden and is linked to several illnesses and

accidents, including cancer. Alcohol use is causally associated with malignancies of the colon, rectum, liver, and female breast in addition to tumors of the upper aerodigestive tract (oral cavity, pharynx, larynx, and oesophagus) [5–6]. Gene expression classification and grouping using microarray or RNA-seq data remains challenging due to the high dimensionality and complexity of the data, making cancer identification difficult despite extensive research. The clinical diagnosis of cancer and identification of tumor-specific markers remain disputed topics, with limitations in interpreting significant signals for classification using machine learning. Although many techniques have been developed to identify cancer-causing genes, their analysis is rarely condensed; The review paper's primary contribution is its investigation of the genes listed below that cause cancer:

- The importance and limits of a dataset analysis for cancer-causing gene identification methodologies are examined in order to improve the classifier accuracy, sensitivity, and specificity in the detection of cancer genes.
- Various machine learning and deep learning techniques, including the performance analysis of cancer-causing gene detection, early diagnosis of cancer-causing genes, hybrid genomics solutions-based DL model, semi-supervised ensemble technique, anisotropic anomaly-based gene detection technique, and hybrid solutions-based DL model are analyzed along with their significance and limitations.

The paper's content is arranged as follows: The literature review of several methods for identifying the genes that cause cancer is presented in Section 2. The methodologies are summarized in Section 3 together with a comparative analysis; the article is concluded in Section 4, and the future outlook is provided in Section 5.

II. LITERATURE REVIEW

In this section, the review has been provided by discussing various technologies in cancer causing gene detection. Additionally, the importance and constraints of these methods for identifying and categorizing cancer genes are discussed. The directions for cancer causing gene detection techniques are shown in figure 1.

This review has been made in five different phases of cancer causing gene detection namely, dataset analysis, Hybrid Radiomics solution-based DL models, Semi-supervised ensemble techniques, Anisotropic Anomaly-based Gene detection techniques, early diagnosis of cancer and performance analysis of cancer causing genes using ML techniques.

A. Review on dataset analysis for cancer causing gene detection

Deng et al.'s [7] bioinformatics study aimed to identify harmful and predictive differentially expressed genes (DEGs) in breast cancer using data from GEO datasets and the Cancer Genome Atlas (TCGA). Utilizing the LIMMA Package in R, they identified 203 up-regulated and 118 down-regulated DEGs. The primary enriched pathways for these genes were the epithelial-to-mesenchymal transition and the mitotic cell cycle. These hub genes may serve as biomarkers for breast cancer. The study suggests a future focus on enhancing the sensitivity and specificity of breast cancer diagnosis. **Shakeel et al. [8]** used enhanced neural and soft computing techniques on lung biomedical data from the EVLIRA data repository, employing bin smoothing normalization, Wolf heuristic, and minimal repetition features to reduce noise, dimensionality, and complexity.

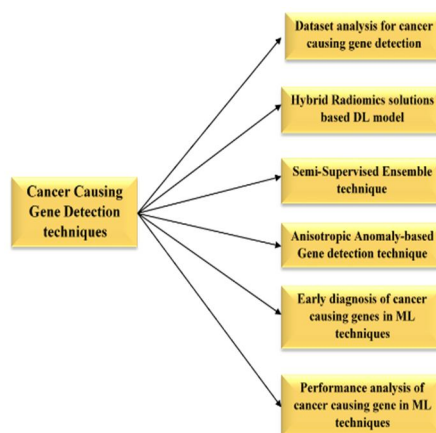


Figure 1: Directions for Cancer causing Gene Detection Techniques

Discrete AdaBoost-optimized ensemble learning then effectively differentiated between normal and abnormal lung characteristics, with future improvements focused on sensor data analysis for more precise early lung cancer prediction. **Falzone et al.** [9] found that new high-throughput technologies in molecular and epigenetic research produce extensive data useful for identifying cancer biomarkers. By analyzing GEO Datasets and TCGA miRNA profiling datasets, they identified miRNAs with potential diagnostic and prognostic significance. Using computational techniques, they determined the functional roles of these miRNAs. However, they concluded that only an integrated analysis of multiple miRNA expression datasets can reliably identify miRNAs for the early diagnosis of oral cancers. **Chen et al.** [10] utilized the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases to identify potential diagnostic and prognostic gene biomarkers for colorectal cancer (CRC). They retrieved RNA-sequencing data from TCGA and microarray data from GEO to analyze CRC gene expression patterns. Their bioinformatics analysis validated the diagnostic and prognostic significance of key genes within the TCGA dataset. However, they recommend future research to validate these findings using tumor samples, clinical data, and other databases. **Song et al.** [11] analyzed differentially expressed genes (DEGs) between hepatocellular carcinoma (HCC) and non-tumor liver tissues using three Gene Expression Omnibus (GEO) databases (GSE46408, GSE65372, and GSE84402). They clarified the functional roles of DEGs through Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment studies. However, they recommend further research with larger sample sizes and experimental data to confirm these findings. This work aims to identify new diagnostic and prognostic biomarkers and therapeutic targets for HCC.

TABLE I: REVIEW ON DATASET ANALYSIS FOR CANCER CAUSING GENE DETECTION

Ref no	Techniques/method	Significance	Limitations
[7]	Bioinformatic analysis	Find harmful and prognostic differentially expressed genes (DEGs) associated with breast cancer.	There is a need to enhance the sensitivity and specificity of breast cancer diagnosis.
[8]	ELVIRA Biomedical Data Set Repository	Successfully differentiated between diseased and normal characteristics using an analysis of the biological lung data.	Need for raising the earlier prediction rates by gathering and reviewing the sensor data for lung cancer prediction
[9]	miRNA expression datasets	Exploration of 11 miRNAs	Need to enhance the diagnosis of oral cancer
[10]	Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) datasets	The top 10 hub genes were found among thousands of candidate genes.	The diagnostic effectiveness and prognostic value of the critical genes were only examined and validated in the TCGA dataset
[11]	Gene Expression Omnibus (GEO) dataset	Investigates the DEG between HCC and non-tumor liver tissues	Needs to be confirmed with larger sample numbers

B. Review on Hybrid Radiomics solutions based DL model in cancer causing gene detection

Xia et al. [12] developed a radiomics technique to identify differentially expressed genes associated with Microvascular Invasion (MVI) in Hepatocellular Carcinoma (HCC) and predict MVI status using preoperative multiphase CT imaging. By extracting radiomic characteristics from tumors and peritumor areas on CT scans, they created five logistic regression models, which were evaluated using internal and external test sets. The best-performing model was combined with clinical-radiologic characteristics to form a hybrid model. Future research should ensure the model's generalizability and assess the biological relevance of the transcriptome sequencing results in cell- or animal-based studies. **Shams et al.** [13] developed a hybrid machine learning model for predicting Hepatocellular Carcinoma (HCC) by combining Naive Bayes, Random Forest, Logistic Regression, and Linear Discriminant Analysis models with a Deep Neural Network as a meta-model. They introduced the Dipper Throated Optimization and Particle Swarm Optimization (DTPSO) technique, combining DTO and PSO algorithms for optimal resolution. Using a dataset of 165 cases and 50 features, they demonstrated that the DTPSO-based model outperformed traditional DTO, PSO, and blending-based models, achieving an accuracy of 98.52% compared to 96.00% for the blending-based model. Future work will focus on further improving prediction accuracy. **Sarkar et al.** [14] proposed a hybrid framework for classifying muscle-invasive bladder cancer (MIBC) vs. non-muscle-invasive bladder cancer (NMIBC), bladder cancer tissue vs. normal tissue, and post-treatment changes (PTC) vs. MIBC. This framework combines statistical machine-learning methods with pre-trained deep neural networks for feature extraction. While the model's diagnostic performance may be overstated due to its retrospective design and small, single-center dataset, it highlights the need for a

multidisciplinary approach in bladder cancer management, with tumor staging being crucial. Future research should use prospective, multi-center datasets for external validation. **Ninomiya et al. [15]** compared homology-based radiomic features (HFs) with traditional wavelet-based radiomic features (WFs) and combined features (HWFs) using training (n = 135) and validation (n = 70) datasets. They constructed 13,824 HFs using Betti numbers to capture topologically invariant morphological aspects of lung cancer. Kaplan-Meier analysis assessed the prognostic potential of HFs, revealing significant differences in survival curves between high- and low-risk patients. However, as the study was based on a single dataset, future research should examine the repeatability of HFs with different scanners, scanning methods, and patient groups to ensure accuracy. **Chang et al. [16]** developed an automated computer-aided detection (CAD) tool for diagnosing pancreatic cancer (PC) using two- and three-dimensional radiomic analysis with machine learning. They retrospectively collected contrast-enhanced CT scans from 546 patients with pancreatic adenocarcinoma and 733 controls with normal pancreas. The CAD tool was tested on a national dataset of 1,477 CT scans from universities across Taiwan, demonstrating high precision and robustness. However, the dataset lacked radiologist reports, preventing comparison with radiologist interpretations. Future research should explore the CAD tool's ability to differentiate between PC and other pancreatic disorders.

TABLE II: REVIEW ON HYBRID RADIOMICS SOLUTIONS BASED DL MODEL IN CANCER CAUSING GENE DETECTION

Ref no.	Techniques/ methods	Significance	Limitations
[12]	Radiomics technique	Identify MVI-associated differentially expressed genes	Need to evaluate the biological applicability of transcriptome sequencing in cell
[13]	Hybrid machine learning model	Predicts the HCC in patients	Need to enhance the accuracy in HCC prediction
[14]	Hybrid framework in DNN	Classifies the cancer tissues	Because the diagnostic approach is retrospective and based on a small dataset from a single center, its performance is overstated.
[15]	Homology-based radiomic features (HFs)	Lung cancer patients' prognosis is predicted.	Only one database is used
[16]	2D and 3D radiomic analysis	Recognizes the PC using CT images enhanced with contrast	NHI dataset did not contain any radiologist reports

C. Review on Semi-Supervised Ensemble techniques for cancer causing gene detection

Alrefai et al [17] in this study, by combining particle swarm intelligence and aggregation ranker filters with the ensemble approach (PSO-ensemble) suggest a structure that can reduce the high dimensionality. After that, to address the low sample size issue, An Adaptive Self-Training Method (ASTM) can be used to the labelled set can be increased by iteratively adding the safest examples from the unlabeled datasets. Empirical research findings demonstrate that ASTM may significantly improve categorization performance. Analyzing the genes that cause cancer is challenging, though. **Wei et al. [18]** proposed a semi-supervised learning technique called CALLR for cell-type annotation, combining supervised learning using sparse logistic regression and unsupervised learning represented by the graph Laplacian matrix. This method iteratively updates annotation labels and cell clusters, formulating the model as an optimization problem and solving it efficiently. Experiments on 10 real datasets show CALLR outperforms other supervised and clustering methods. However, its ability to detect unusual cell types may be limited by initial cell identification challenges and small sample sizes. **Hamamoto et al. [19]** developed GraphPheno, a partially-supervised approach using graph auto-encoders to extract deep features from protein sequences without feature engineering. It also considers the topological characteristics of the protein-protein interaction network to predict connections between human genes, proteins, and abnormal phenotypes. Cross-validation and tests on separate datasets demonstrate GraphPheno's adequate prediction performance. In the future, GraphPheno will help identify unique gene-phenotype associations and serve as a valuable tool for studying diseases and phenotypes. **Prathik et al [20]** a deep reinforcement learning (DRL) model that may be used to efficiently analyze gene expression data and determine the kind of cancer. Analyzing the model for predicting the variant types of cancer involves using the database of gene expression. Comparing the proposed DRL model to other existing models, the simulation results demonstrate that it can accurately predict the kind of cancer with a 97.8% accuracy rate. However, in the future, there is a need to increase the prediction accuracy. **Wang et al. [21]** introduced HFHLMMDA, a unique prediction approach that uses hypergraph learning with high-dimensionality features to reveal relationships between diseases and miRNAs. This method combines miRNA functional similarity and disease semantic resemblance to create a high-dimensional feature vector. Using the K-Nearest-Neighbor (KNN) approach, a hypergraph is constructed to represent complex miRNA-disease

interactions by linking each pair with its k most relevant neighbors. The final stage involves learning a projection matrix to determine unknown miRNA-disease association scores. Although ensuring an optimal hypergraph structure for all applications is challenging, future research will focus on optimizing hypergraph structures, leading to a dynamic learning approach.

TABLE III: REVIEW ON SEMI-SUPERVISED ENSEMBLE TECHNIQUE FOR CANCER CAUSING GENE DETECTION

Ref no	Techniques/methods	Significance	Limitations
[17]	Aggregation ranker-filters and particle swarm intelligence with the ensemble method (PSO-ensemble)	Minimize the high dimensionality	It is hard to analyze the cancer causing genes
[18]	Sparse logistic regression and unsupervised learning	Enumerate the cell types	It might not be able to detect unusual cell types.
[19]	Similarity-preserving method and a sparse regularization term	Effectively incorporates labeled and unlabeled components into the co-selection procedure	Need to consider co-selection for multi-label classification and regression issues
[20]	Deep learning-based model (Deep- Noise)	Separate the true biological signals from the interfering technological noises	The test dataset's labels are not accessible
[21]	HFHLMDA	Reveals the relationship between illnesses and miRNAs	It is challenging to ensure that the created hypergraph structure is perfect and suitable for every application.

D. Review on Anisotropic Anomaly-based Gene detection techniques

Gomathi et al. [22] proposed an unsupervised anisotropic-feature transformation technique for analyzing breast thermal images, classifying them into malignant, benign, and standard categories. This technique uses screening as a diagnostic tool to enhance breast cancer detection accuracy. The results indicate that this method is promising for identifying cancer-related deviations in women. The study highlights the importance of further promoting this technique and conducting more trials. Additionally, to reduce misclassification, future developments should focus on incorporating deep learning methods. **Kabir et al [23]** proposed an algorithm for automatically identifying and categorizing brain tumors. Five phases made up the proposed algorithm. An anisotropic diffusion filter and principle component-based grayscale conversion are first used to preprocess the raw MRI image in order to eliminate unwanted artefacts. The picture contrast is then enhanced by applying the contrast-limited adaptive histogram equalization (CLAHE). Next, multivariant thres holding and the Chan-Vese algorithm are used to segment the tumor. However, in the future, Deep learning-based segmentation and classification can more effectively improve the multiscale classification. **Jia et al. [24]** proposed three hybrid algorithms combining the thermal exchange optimization (TEO) method with the seagull optimization algorithm (SOA) to address feature selection challenges. The first algorithm uses a roulette wheel to alternate between two updating algorithms. The second algorithm incorporates the TEO method after SOA iterations for further optimization. The third algorithm modifies the SOA's seagull attack mode using the TEO's heat exchange formula to enhance exploitation capacity. Although the approach was not tested on a real dataset, the modified method shows improved feature selection. Further research is recommended to simplify the optimization algorithms and apply this method to other sectors for more precise feature selection.

Bourouis et al. [25] developed a computer-aided diagnosis approach for detecting abnormalities in breast ultrasound images using a meta-heuristic method to adjust neural network parameters. This novel method combines the grey wolf optimization (GWO) algorithm with a wavelet neural network (WNN). Preprocessing of breast ultrasound images involves interference-based despeckling, anisotropic diffusion, and a sigmoid filter. The region of interest is extracted using an automatic segmentation algorithm, followed by calculating morphological and texture features. Classification is then performed using the GWO-tuned WNN. The technique was tested on 346 ultrasound images, and its efficacy was evaluated using the receiver operating characteristic (ROC) curve and confusion matrix. Future research will explore other meta-heuristic approaches, such as ant lion optimization and cat swarm optimization, to further improve classification accuracy. **Ijaz et al. [26]** proposed a cervical cancer prediction model (CCPM) that uses risk indicators to predict cervical cancer early. The CCPM removes outliers by employing methods like isolation forest (iForest) and density-based spatial clustering of applications with noise (DBSCAN) for unsupervised anomaly detection. However, this approach, tested on a single dataset, uses more memory and operates more slowly compared to data balancing using Random Forest (RF).

TABLE IV: REVIEW ON ANISOTROPIC ANOMALY-BASED GENE DETECTION TECHNIQUE

Ref no	Techniques/methods	Significance	limitations
[22]	Unsupervised anisotropic-feature transformation	Identifying cancer to create a legal pathway for promoting effective	Incorrect classification ratio
[23]	Automatic brain tumor detection and classification algorithm	Effectively identify and categorize the brain tumor	Need to improve the multiscale classification
[24]	Thermal exchange optimization (TEO) method and the seagull optimization algorithm (SOA)	Effectively handle feature selection challenges	Feature selection for the real data set is not done.
[25]	Anisotropic diffusion	Preprocessing the images from breast ultrasonography (US)	Need to enhance the classification accuracy
[26]	Cervical cancer prediction model (CCPM)	Predicts cervical cancer early	Only one dataset is used

E. Review on early diagnosis of cancer causing genes using ML techniques

Dhahri et al. [27] used genetic programming and machine learning to develop a technology for accurately distinguishing between benign and malignant breast tumors. The study optimized the learning algorithm by selecting the best features and ideal parameter values for the classifiers using genetic programming. The effectiveness of the proposed technique was evaluated using ROC curves, sensitivity, specificity, precision, and accuracy. The study demonstrates that genetic programming, combined with feature preprocessing and classifier algorithms, can automatically identify the optimal model. This approach could be extended to other cancer predictions in the future.

Begum et al. [28] proposed a model combining Fuzzy Preference-Based Rough Set (FPRS) with Support Vector Machine (SVM) to predict cancer biomarkers using biological and gene expression databases. This model evaluates the importance of genes based on specific criteria, using Rough Set theory to identify crucial and unnecessary factors. The method was tested on a biological dataset and four gene expression databases, successfully identifying informative genes and biological indicators. However, it is difficult for users to input such information into the program, necessitating an algorithm for this task. Additionally, the approach faces issues with fuzzy approximation consistency and SVM's lack of transparency. Future research should use the imperialist competition algorithm to further assess the model's effectiveness. **Abdar et al. [29]** proposed a nested ensemble approach using classifier arranging and voting to differentiate between benign and malignant breast tumors. Each hierarchical ensemble classifier includes "Classifiers" and "Meta Classifiers," where a meta classifier can integrate multiple distinct classification methods. For this study, two-layer nested ensemble classifiers were developed, with meta classifiers using two or three different classification methods. The Wisconsin Diagnostic Breast Cancer (WDBC)

dataset and K-fold Cross Validation were used to evaluate the model. Future research should explore different ensemble modeling approaches for various cancers and diseases to enhance diagnostic effectiveness. Additionally, the complexity and size of medical databases have rapidly increased. **Jeyaraj et al. [30]** developed a deep learning technique for an autonomous cancer diagnostic system using a partitioned convolutional neural network (CNN). They analyzed hyperspectral medical images for oral cancer, utilizing the stochastic neighbor embedding method to visually display features. The partitioned deep CNN outperformed traditional classification methods like SVM and DBN, achieving a specificity of 0.98, sensitivity of 0.94, and accuracy of 94.5%. For 500 training patterns, the system accurately categorized malignant tumors, benign tumors, and normal tissue with 94.5% accuracy. For 100 image datasets, it achieved a categorizing accuracy of 91.4%, with a sensitivity of 0.94 and specificity of 0.91. Future work should focus on enhancing the automated cancer detection system. **Ganggayah et al. [31]** utilized machine learning techniques to examine prognostic indicators for breast cancer survival. Among the methods tested, the random forest technique showed slightly better accuracy compared to others, though the accuracies of all algorithms were fairly similar. The study used a large dataset from the University Malaya Medical Centre, containing diagnostic data for 8,066 breast cancer patients from 1993 to 2016, including 23 predictor factors and the patient's survival status. Prediction models were developed using logistic regression, decision trees, random forests, neural networks, extreme boost, and support vector machines to identify key prognostic factors. Advanced modeling with random forests was further refined by clustering the dataset based on the immunohistochemistry-determined receptor status of patients. Decision trees were then validated using survival analysis. Future research should focus on enhancing medical predictive applications for various disease prognoses.

TABLE V: REVIEW ON EARLY DIAGNOSIS OF CANCER CAUSING GENES IN ML TECHNIQUES

Ref no.	Techniques/ methods	Significance	Limitations
[27]	Genetic programming and machine learning techniques	Categorize the benign and malignant breast tumors with accuracy	Need to reduce overfitting problems
[28]	Fuzzy Preference-Based Rough Set (FPRS) combined with SVM	Predicting cancer biomarkers using gene expression and biology databases	In SVM, there is a lack of results transparency
[29]	Nested ensemble approach	Separates benign breast tumors from malignant malignancies	Need to reduce the complexity and amount of medical databases
[30]	Deep learning based on a partitioned convolution neural network	Predicted whether the tumor was benign or malignant.	There is a need to enhance the automated cancer detection system.
[31]	Machine learning techniques	Identifying and presenting major prognostic factors of breast cancer survival rates.	Need to enhance the medical predictive applications for analyzing the prognosis of various diseases.

F. Review on performance analysis of cancer causing gene detection techniques

Wu et al. [32] utilized machine learning techniques to analyze gene expression data for classifying patients with triple-negative breast cancer (TNBC) and those without it. They examined RNA-Sequence data from The Cancer Genome Atlas, including 110 TNBC and 992 non-TNBC tumor samples, to select genes for constructing and validating classification models. Four different model designs—Support Vector Machines, K-nearest Neighbors, Naive Bayes, and Decision Trees—were investigated using selected variables at various threshold levels to classify breast cancer types. Their approach allowed for functional evaluation and validation across separate gene expression databases. Future studies should focus on further assessing the efficacy of ML algorithms in classifying TNBC and non-TNBC subtypes, optimizing classification features, and integrating radiomics and genomic data for enhanced analysis. **Khalifa et al. [33]** introduced an advanced deep learning approach, BPSO-DT, combining Particle Swarm Optimization in Binary with a decision tree (DT) and convolutional neural network (CNN), to classify various types of cancer based on tumor RNA sequence (RNA-Seq) gene expression data. The strategy involves preprocessing with BPSO-DT to select optimal features from high-dimensional RNA-Seq data and converting it into 2D images. Data augmentation expands the initial dataset to 10,430 samples using strategies that minimize feature alteration. The deep CNN architecture is then employed to train the model, improving accuracy and addressing overfitting. Evaluation metrics such as recall, precision, and F1 score demonstrate an overall testing accuracy of 96.90%.

Future research could extend this approach to categorize 33 types of cancer by exploring diverse deep learning architectures like AlexNet, Vgg-16, and GoogleNet, contingent upon the availability of suitable datasets. **Lai et al. [34]** developed a deep neural network (DNN) using various clinical and gene expression data sources to accurately predict overall survival in patients with non-small cell lung cancer (NSCLC). They used microarray data from 614 patients to classify them into biomarker- and biomarker+ groups based on seven well-known NSCLC biomarkers. Prognosis relevance values (PRV) were calculated using a systems biology approach to identify eight additional prognostic gene biomarkers. Combining these 15 biomarkers with clinical data, they created an integrated DNN using bimodal learning to predict 5-year survival with high accuracy. Future work should focus on reducing the computational time required for predicting cancer survival. **Basavegowda et al. [35]** developed a deep feedforward classification algorithm to analyze microarray cancer data for future diagnosis. They designed a 7-layer deep neural network, addressing small sample sizes and dimensionality issues using principal component analysis. The method was validated on eight common microarray cancer datasets, with feature values scaled using the Min-Max method. Optimization was achieved through the adaptive moment estimation, and binary cross-entropy was used to measure loss. The performance metrics included classification precision, recall, f-measure, log-loss, receiver operating characteristic curve, and confusion matrix. Future work should expand this strategy to multi-class microarray cancer datasets and aim to improve classification accuracy for lower-ranking binary datasets. **Vaka et al. [36]** introduced a Deep Neural Network Support (DNNS) method for detecting breast cancer using machine learning algorithms. They conducted an experimental analysis on a dataset, demonstrating that their method is more accurate and effective than existing methods. The segmentation process involved calculating and comparing performance metrics such as precision, recall, sensitivity, accuracy, F-measure, and rank sum against current methods like fuzzy C-means (FCM) and thresholding. A normalization procedure was used to enhance performance, efficiency, and image quality. Experimental results showed that DNNS significantly outperforms current approaches. While promising, future work should focus on improving the prediction accuracy for breast cancer detection.classifiers' methods have high correctness ratings of 94.07%,

sensitivity ratings of 92.05%, and specificity ratings of 94.07%. However, in the future, there is a need to enhance the classifiers' accuracy in prediction.

TABLE VI: REVIEW ON PERFORMANCE ANALYSIS OF CANCER CAUSING GENE IN ML TECHNIQUES

Ref no.	Techniques/ methods	Significance	Limitations
[32]	SVM algorithms	Reduced misclassification errors and a more reliable ability to distinguish between triple negative and non-triple negative breast cancer.	The efficiency of machine learning algorithms in identifying TNBC and non-TNBC subtypes must be ascertained.
[33]	Optimized deep learning strategy	Using information on tumor RNA sequence (RNA-Seq) gene expression, classify different types of cancer.	Need to introduce more datasets to find the different types of cancer
[34]	Deep neural network (DNN)	Within five years, predicted the patients' survival status	Need to enhance the computational time in predicting the cancer survival of patients.
[35]	Deep feedforward neural network method	Classifies the binary class microarray datasets.	Less classification accuracy
[36]	Deep Neural Network Support value (DNNS)	Finds the breast cancer	Need to enhance the prediction accuracy

III. RESULT AND SUMMARY

The Machine Learning (ML) and deep learning (DL) based detection techniques for cancer causing genes have been analyzed in various directions such as dataset analysis, Hybrid Radiomics solutions-based DL model, semi-supervised ensemble learning Anisotropic Anomaly-based Gene detection techniques in cancer, early diagnosis of cancer causing genes in ML techniques, and performance analysis of cancer causing gene in ML techniques. The analyzed summary is given as follows:

- Dataset analysis of various methods such as bioinformatic analysis, biomarker mRNAsi, Gene Expression Omnibus (GEO) are performed, However, these analyses are needed to enhance the sensitivity and specificity of cancer diagnosis, need to raise the earlier prediction rates for cancer, and more fundamental biological research is required.
- Hybrid Radiomics solutions based DL models in cancer causing gene detection methods such as Radiomics analysis, Hybrid machine learning model are analyzed. However, it is only limited to predicting a person's receptor status and proliferation rate, need to enhance the accuracy in HCC prediction, diagnostic performance in cancer is inflated, and only one dataset is used for cancer gene prediction.
- Many semi-supervised ensemble learning for cancer causing genes detection such as particle swarm intelligence and aggregation ranker-filters using the ensemble approach are analyzed, However, it is hard to analyze cancer causing genes, not able to detect unusual cell types in cancer, the test dataset's labels are not accessible and the prediction accuracy is very low
- Various Anisotropic Anomaly-based Gene detection techniques in cancer such as unsupervised anisotropic-feature transformation, automatic brain tumor detection, are analyzed, however it is restricted to the examination of tissues dyed with exogenous fluorescent dyes in vitro. failure to detect the pathology, and feature selection for the real data set is not done in the Anisotropic Anomaly-based Gene detection technique.
- Early diagnoses of cancer causing genes in ML techniques such as CNN, genetic programming in ML are analyzed, However, only the metabolite level was included in the data, biomarkers for earlier cancer detection are poor, the size of the two-class dataset is very small in GP, and in SVM there is a lack of results transparency.
- Performance analysis of various cancer causing gene detection techniques in machine learning such as SVM algorithms, optimized deep learning strategy are analyzed, However, there is only an 8% improvement is attained in the identification of cancer causing genes, there is a significant effect on the therapeutic care of individuals with this disease, need to enhance the conventional feature extraction network and the classification accuracy is low.

A. Comparison of performance of various Machine Learning and Deep Learning techniques for cancer causing gene detection

The comparison examination of various Machine Learning and Deep Learning techniques for cancer causing gene detection is discussed in this section.

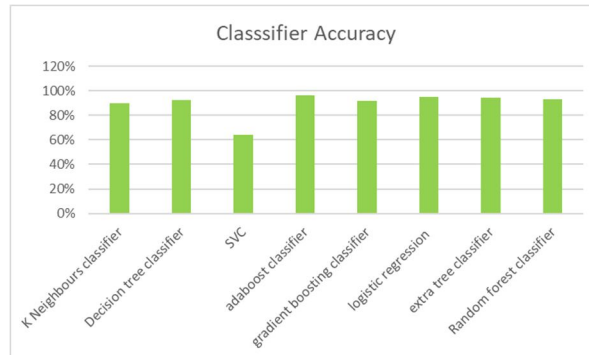


Figure 2: Comparison of various classifiers for classifier accuracy

Figure 2 shows the classifier accuracy of various classifiers such as the K-neighbours classifier, Decision tree classifier, SVC, AdaBoost, logistic regression, additional tree, random forest, and gradient boosting classifiers are among the classifiers. From the comparison, it is observed that AdaBoost classifier attains the maximum classifier accuracy of 97.82% and SVC attains the minimum classification accuracy of 64%. Since the classifier accuracy in cancer gene detection is reduced due to only limited of predicting a person's receptor status and proliferation rate.

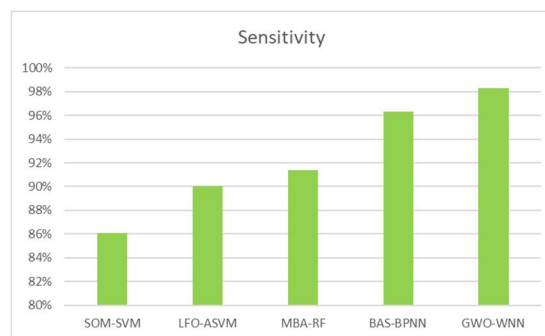


Figure 3: Comparison of sensitivity in predicting cancer genes

Figure 3 represents the comparison of sensitivity in detecting cancer causing genes by using some techniques such as SOM-SVM, LFOA-SVM, MBA-RF, BAS-BPNN and GWO-WNN. From the comparison, it is noted that for SOM-SVM approach the sensitivity attains 86% which is small while in contrast to alternative techniques. The sensitivity for detecting the cancer causing genes in GWO-WNN attains its maximum at 98.02%. Meanwhile, the sensitivity of cancer diagnosis is poor in which the sensitivity of SOM-SVM is 86 % which is the minimum compared to other existing approaches.

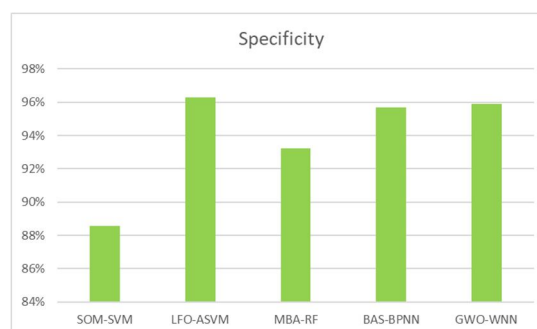


Figure 4: Comparison of specificity in predicting cancer genes

Figure 4 shows the specificity of various cancer gene detection approaches, including SOM-SVM, LFOA-SVM, MBA-RF, BAS-BPNN, and GWO-WNN. From the comparison, it can be seen that the specificity for the SOM-SVM strategy is only 88.54%, which is low when compared to other methods. The LFOA-SVM reaches its greatest specificity for detecting the genes that cause cancer at 96.3%. Moreover, the specificity of cancer diagnosis is poor in which the specificity of SOM-SVM is 88.54 % which is the minimum compared to other existing approaches.

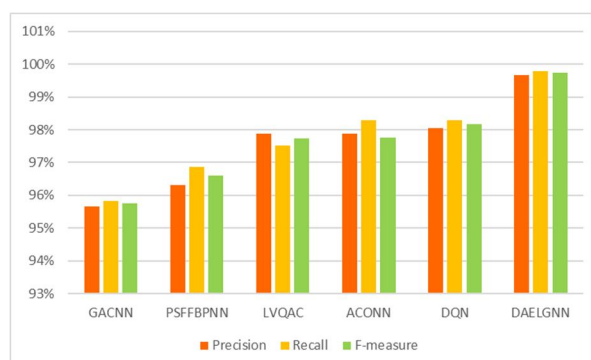


Figure 5: Comparison of precision, recall, and F-measure

Figure 5 represents the comparison of the precision, recall, and F-measure in detecting the cancer causing genes. Comparing these with Genetic maximised optimized constructive neural networks (GACNN), Particle swarm feed-forward back-propagation neural networks (PSFFBPNN), Learning vector quantization optimized with ant colony (LVQAC), Ant colony optimization neural networks (ACONN), Deep Q-networks (DQN), and Discrete AdaBoost optimized ensemble learning generalized neural networks (DAELGNN). From the comparison it is noted that precision recall and F-measure for DAELGNN is maximum at 99.67%, 99.78% and 99.73% respectively. Here, GACNN attains its minimum value for precision, recall and F-measure at 95.67%, 95.82%, and 95.75% respectively. However, compared to other methods now in use, the accuracy, recall, and F-measure of GACNN for identifying cancer genes are poor.

Overall, the comparison analysis of various Machine Learning and Deep Learning techniques for cancer causing gene detection is presented. Here, the classifier accuracy of various classifiers such as the K-neighbours classifier, there is a comparison of decision tree, SVC, AdaBoost, gradient boosting, logistic regression, extra tree, and random forest classifiers. Moreover, the sensitivity and specificity in detecting cancer causing genes by using some techniques such as SOM-SVM, LFOA-SVM, MBA-RF, BAS-BPNN and GWO-WNN. However, the sensitivity and specificity of cancer diagnosis is poor, since it is minimum due to the various limitation associated with them. Additionally, when compared to other methods currently in use, the accuracy, recall, and F-measure for identifying the cancer genes in GACNN are poor.

IV. CONCLUSION

This research reviews various deep learning and machine learning methods for detecting cancer-causing genes, including dataset analysis, hybrid radiomics solutions, semi-supervised ensemble learning, and anisotropic anomaly-based detection. The study addresses the limitations of previous methods, such as inappropriate feature extraction, imprecise prediction, poor diagnosis, and detecting complexity, by employing advanced ML techniques. It highlights the challenges in extracting and predicting semi-quantitative features for tumor suppressor genes and proto-oncogenes. The proposed approach uses hybrid machine learning to effectively differentiate between benign and malignant tumors, enhancing detection rates, feature extraction, and diagnosis efficiency for early cancer diagnosis.

FUTURE PERSPECTIVE

Future research can develop the following suggestions to address the limitations identified:

Hybrid Machine Learning Algorithms: Improve the sensitivity and specificity of breast cancer diagnosis, enhance oral cancer diagnosis, and increase the likelihood of early lung cancer detection using hybrid machine learning algorithms.

Hybrid Radiomics Solutions-Based Deep Learning Model: Develop this model to increase the accuracy of HCC prediction, enhance receptor predicting status and proliferation rates, and improve diagnostic performance by extracting potential features from medical images for customized treatment.

Semi-Supervised Ensemble Technique: Enhance this technique to analyze cancer-causing genes' hardness, detect unusual cell types, and increase prediction accuracy using gene expression data, thus providing insights into precise and successful cancer therapy.

Anisotropic Anomaly-Based Gene Detection Technique: Improve this technique for better classification ratios in cancer gene detection, enabling the identification of tumor suppressor genes and proto-oncogenes through the Radial Basis Random Forest algorithm for effective cancer-causing gene detection.

Early Diagnosis Through ML and DL Techniques: Enhance these techniques to detect cancer-causing genes earlier, using methods like the Enhanced Grading-Based Gradient Boosting method to assign cancer stages based on DNA damage for early diagnosis of malignant tumors.

Efficient Machine Learning Techniques for Gene Detection: Develop efficient techniques to improve conventional feature extraction networks, classification accuracy, and classifier performance, using a Statistical Prognostic Model to differentiate between benign and malignant tumors and integrated signal-to-noise ratio-based ranking to prevent loss of tumor suppressor and proto-oncogene functions.

REFERENCES

- [1] Saba, Tanzila. "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges." *Journal of infection and public health* 13.9 (2020): 1274-1289.
- [2] Han, Yi, et al. "DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies." *Nucleic acids research* 47.8 (2019): e45-e45.
- [3] de Martel, Catherine, et al. "Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis." *The Lancet global health* 8.2 (2020): e180-e190.
- [4] Ali, Engy A., and Mariam Raafat. "Relationship of mammographic densities to breast cancer risk." *Egyptian Journal of Radiology and Nuclear Medicine* 52 (2021): 1-5.
- [5] Rumgay, Harriet, et al. "Global burden of cancer in 2020 attributable to alcohol consumption a population-based study." *The Lancet Oncology* 22.8 (2021): 1071-1080.
- [6] Sun, Qiuyu, et al. "Alcohol consumption by beverage type and risk of breast cancer: a dose-response meta-analysis of prospective cohort studies." *Alcohol and Alcoholism* 55.3 (2020): 246-253.
- [7] Deng, Jun-Li, Yun-hua Xu, and Guo Wang. "Identification of potential crucial genes and key pathways in breast cancer using bioinformatic analysis." *Frontiers in genetics* 10 (2019): 695.
- [8] Shakeel, P. Mohamed, et al. "RETRACTED ARTICLE: Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks." *Neural Computing and Applications* 32.3 (2020): 777-790.
- [9] Falzone, Luca, et al. "Identification of novel MicroRNAs and their diagnostic and prognostic significance in oral cancer." *Cancers* 11.5 (2019): 610.
- [10] Chen, Linbo, et al. "Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis." *Gene* 692 (2019): 119-125.
- [11] Song, Xiudao, et al. "Identification of potential hub genes related to the progression and prognosis of hepatocellular carcinoma through integrated bioinformatics analysis." *Oncology reports* 43.1 (2020): 133-146.
- [12] Xia, Tian-yi, et al. "Predicting microvascular invasion in hepatocellular carcinoma using CT-based radiomics model." *Radiology* 307.4 (2023): e222729.
- [13] Shams, Mahmoud Y., et al. "A hybrid dipper throated optimization algorithm and particle swarm optimization (DTPSO) model for hepatocellular carcinoma (HCC) prediction." *BiomedicalSignal Processing and Control* 85 (2023): 104908.
- [14] Sarkar, Suryadip, et al. "Performing automatic identification and staging of urothelial carcinoma in bladder cancer patients using a hybrid deep-machine learning approach." *Cancers* 15.6 (2023): 1673.
- [15] Ninomiya, Kenta, and Hidetaka Arimura. "Homological radiomics analysis for prognostic prediction in lung cancer patients." *Physica Medica* 69 (2020): 90-100.
- [16] Chang, Dawei, et al. "Detection of pancreatic cancer with two-and three-dimensional radiomic analysis in a nationwide population-based real-world dataset." *BMC cancer* 23.1 (2023): 58.
- [17] Alrefai, Nashat, and Othman Ibrahim. "Semi-supervised ensemble learning for expanding the low sample size of microarray dataset." *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. IEEE, 2021.
- [18] Wei, Ziyang, and Shuqin Zhang. "CALLR: a semi-supervised cell-type annotation method for single-cell RNA sequencing data." *Bioinformatics* 37.Supplement_1 (2021): i51-i58.
- [19] Hamamoto, Ryuji, et al. "Epigenetics analysis and integrated analysis of multiomics data, including epigenetic data, using artificial intelligence in the era of precision medicine." *Biomolecules* 10.1 (2019): 62.

- [20] Prathik, A., et al. "Prediction of carcinoma cancer type using deep reinforcement learning technique from gene expression data." *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2021*. Singapore: Springer Nature Singapore, 2022. 541-552
- [21] Wang, Yu-Tian, et al. "MiRNA-disease association prediction via hypergraph learning based on high-dimensionality features." *BMC Medical Informatics and Decision Making* 21 (2021): 1-13.
- [22] Gomathi, P., C. Muniraj, and P. S. Periasamy. "RETRACTED: Breast thermography based unsupervised anisotropic-feature transformation method for automatic breast cancer detection." (2020): 103137.
- [23] Kabir, Md Ahasan. "Automatic brain tumor detection and feature extraction from mriimage." *GSI* 8.4 (2020): 695-711
- [24] Jia, Heming, Zhikai Xing, and Wenlong Song. "A new hybrid seagull optimization algorithm for feature selection." *IEEE access* 7 (2019): 49614-49631.
- [25] Bourouis, Sami, et al. "Meta-heuristic algorithm-tuned neural network for breast cancer diagnosis using ultrasound images." *Frontiers in Oncology* 12 (2022): 834028.
- [26] Ijaz, Muhammad Fazal, Muhammad Attique, and Youngdoo Son. "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods." *Sensors* 20.10 (2020): 2809.
- [27] Dhahri, Habib, et al. "Automated breast cancer diagnosis based on machine learning algorithms." *Journal of healthcare engineering* 2019.1 (2019): 4253641
- [28] Begum, Shemim, et al. "Identification of biomarker on biological and gene expression data using fuzzy preference based rough set." *Journal of Intelligent Systems* 30.1 (2020): 130-141.
- [29] Abdar, Moloud, et al. "A new nested ensemble technique for automated diagnosis of breast cancer." *Pattern Recognition Letters* 132 (2020): 123-131.
- [30] Jeyaraj, Pandia Rajan, and Edward Rajan Samuel Nadar. "Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm." *Journal of cancer research and clinical oncology* 145 (2019): 829-837.
- [31] Ganggayah, Mogana Darshini, et al. "Predicting factors for survival of breast cancer patients using machine learning techniques." *BMC medical informatics and decision making* 19 (2019): 1-17.
- [32] Wu, Jiande, and Chindo Hicks. "Breast cancer type classification using machine learning." *Journal of personalized medicine* 11.2 (2021): 61.
- [33] Khalifa, Nour Eldeen M., et al. "Artificial intelligence technique for gene expression by tumor RNA-Seq data: a novel optimized deep learning approach." *IEEE Access* 8 (2020): 22874-22883
- [34] Lai, Yu-Heng, et al. "Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning." *Scientific reports* 10.1 (2020): 4679.
- [35] Basavegowda, Hema Shekar, and Guesh Dagnew. "Deep learning approach for microarray cancer data classification." *CAAI Transactions on Intelligence Technology* 5.1 (2020): 22-33.
- [36] Vaka, Anji Reddy, Badal Soni, and Sudheer Reddy. "Breast cancer detection by leveraging Machine Learning." *Ict Express* 6.4 (2020): 320-324.