IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 May 2024, accepted 3 June 2024, date of publication 10 June 2024, date of current version 18 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3411633

RESEARCH ARTICLE

Cancer Disease Prediction Using Integrated Smart Data Augmentation and Capsule Neural Network

U. RAVINDRAN[1] **AND C. GUNAVATHI**[2]
[1]School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore 632014, India
[2]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, India

Corresponding author: C. Gunavathi (gunavathi.cm@vit.ac.in)

This work was supported by Vellore Institute of Technology, Vellore, India.

ABSTRACT Cancer accounts for a considerable portion of the illnesses that cause early human death worldwide, and this trend is expected to worsen in the coming years. Therefore, timely and precise identification would be tremendously helpful for cancer patients. Gene expression datasets are commonly utilized for disease identification, particularly in cancer therapy. Deep learning (DL) has become a popular technique in healthcare because of the abundance of computational capacity. The gene expression data samples for five types of cancer disease and healthy samples are collected, but the samples in the gene data are insufficient to fulfill the deep learning requirements. To increase the training sample size, data augmentation is frequently used. The main objective of this research is the diagnosis and classification of different types of cancer. In this research, correlation-centered feature selection and reduction are used to select the most relevant features from the large volume of gene information. The proposed method is a smart data augmentation process with the CapsNet (Capsule Neural Network) method for the accurate prediction and classification of cancer diseases. The proposed augmentation strategy integrates Uniform Distributive Augmentation (UDA) and a Wasserstein-Generative Adversarial Network (W-GAN). The synthetic data samples are generated using uniform distribution and Wasserstein distance, and the newly evolved datasets are employed to train CapsNet. Then, the practical outcome of the integrated smart data augmentation with CapsNet is compared with other DL methods. As a result, the proposed method enhances the classification accuracy, precision, and recall values (>98%) and reduces the error rate.

INDEX TERMS CapsNet, UDA, W-GAN, cancer disease, gene expression data.

I. INTRODUCTION

Cancer is a deadly disease that affects people of all ages and from all geographical places. Figure 1 shows cancer across the globe based on the International Agency for Research on Cancer (IARC) [1]. Cancer is the leading cause of death worldwide, with approximately 10 million deaths. One in six deaths globally is caused by cancer. According to the latest cancer case statistics, female breast cancer has surpassed lung cancer as the most common type of cancer diagnosed, accounting for around 11.7% of all new cases, followed by lung cancer at 11.4%, colorectal cancer at 10.4%, prostate cancer at 7.3%, and stomach cancer at 5.6%. According to cancer death statistics, lung cancer continues to be the leading cause of cancer death at roughly 18%, followed by colorectal cancer at 9.4%, liver cancer at 8.3%, stomach cancer at 7.7%, and female breast cancer at 6.9% [2]. As a result, statistical data reveals the importance of controlled research procedures that utilize advanced technologies like deep learning (DL) to classify and predict cancer disease.

Cancer affects people in a variety of ways. Thus, the situation must be closely monitored. The amount of cancer data that researchers can access has increased dramatically because of technological improvements. However, extracting correct inferences from such data is a challenging task. This can be achieved through the application of artificial intelligence (AI) methods. AI approaches can now be used to investigate patterns in large and complex datasets. The amount of data that deep learning models can analyze is limitless [3]. Deep learning (DL) algorithms are widely employed

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

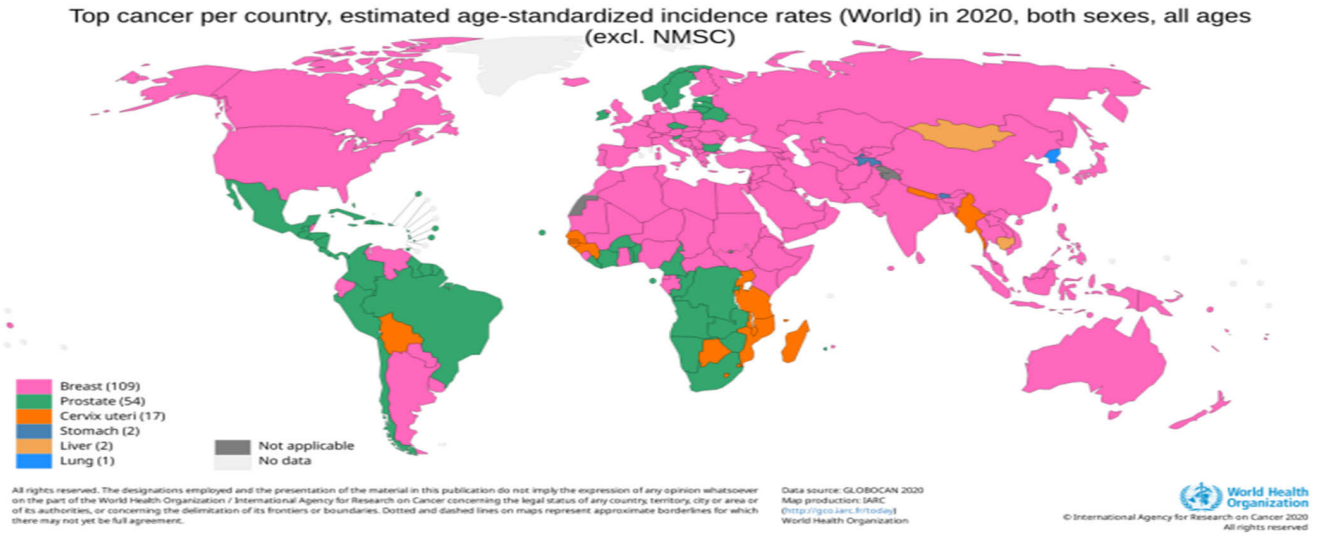

FIGURE 1. Cancer across the globe based on international agency for research on cancer (IARC).

in diagnostic genomics. DL approaches have been applied to the classification, prediction, and conditional estimate of cancer disease, strengthening decision-making. DL methods like Feed Forward Neural Network, Convolutional Neural Network, Recurrent Neural Network, and Auto encoder were analyzed to predict the cancer disease using gene expression data [4].

Molecular genetics research has been essential to our understanding of this disease for the last two decades. Early identification of cancer disease must be a top priority for oncology researchers. The microarray gene expression cancer data (MGECD) is more appropriate for cancer classification, which will ultimately support qualitative research. The genome database consists of a larger number of features in gene expressions and fewer data samples. To completely understand the cause of cancer and forecast possible samples, compare the genes expressed in abnormal malignant cells to those expressed in normal cells. It provides a reasonable assessment of gene activity within a specific tissue. A wide variety of features and expression patterns must be taken into account when creating a cancer prediction model. Consequently, this problem affects the accuracy of any model and increases the computation time [5].

Two main techniques have been proposed to minimize the dimensionality issues in the gene expressions as well as to address their subsequent problems. First, we use data augmentation techniques to enhance the new data samples based on low-dimensional attributes (identifiers). Second, the selection of features eliminates the non-relevant reliant traits in favor of more pertinent discriminating ones. Previous investigations have shown that microarrays are clearly performing a critical role in genomic research, with the ultimate goal of developing methodologies to analyze cancer disease and estimate oncologic health outcomes. To improve the efficiency of any DL methodology, the most reliable way to increase the sample size in the dataset is through data augmentation. The MGECD has proven to be a reliable source in the field of cancer research. Sample sets of gene-expressed data are primarily available for study in the accessible realm of genomics. Both the processed and unprocessed versions of such datasets are readily accessible. In addition, structural datasets for genetic expression can be commonly found in many data repositories. However, significant issues with these datasets must be addressed if the desired results are to be achieved. The dimensionality burden in MGECD is the most challenging issue to address. Compared to the number of attributes, the sample size used in testing is quite minimal.

Thus, we introduced two-level segmentation techniques for data augmentation purposes, where the gene dataset is augmented via a UDA strategy and further augmented using Wasserstein-GAN. The Wasserstein distance (WD) was one of the crucial measures for comparing probabilistic ranges. It is also referred to as Earthmover's Distance and is primarily utilized in the realm of the ideal distribution of data patterns and their relative analysis. In fact, even when two distinct distributions are positioned in a low-dimensional space where they don't overlap, the WD can still provide a coherent and consistent approximation of the distance between them. Consequently, only the WD delivers a continuous metric that is very beneficial for steady learning. Consequently, it's primed to address the instability problem that affects standard GANs. WD is a loss function to minimize the distance between synthetic and actual data. Because gene expression is a complicated and difficult problem to solve, such genetic data are either directly or indirectly related to one another, making the prediction process highly complex and typically necessitating an effective and robust feature selection approach. Since Pearson's correlative coefficient (CC) evaluates the degree

of correlation among variables of interest, we incorporated Pearson's CC in this research for feature selection.

Neural networks (NNs), which are based on Brian's cognitive system, have been developed and are currently being utilized for a variety of applications. This system is fed a range of inputs and trained to produce the desired output. Given a set of input parameters and a target, the NN may be a non-linear process. Even in the absence of human supervision, a neural network (NN) can be trained and learned on large amounts of complex raw data on its own. For instance, the Deep NN method is utilized for cancer diagnosis and prediction using MGECD. It is crucial to determine which genes significantly influence the occurrence of cancer. Identifying relevant cancer genes is the only way that prevention and evaluation can progress. Convolutional Neural Network (CNN) algorithms have proven efficient in a number of investigations and are primarily utilized to retrieve facts related to symptoms connected with cancer. These methods find the corresponding patterns based on the expression patterns that are presented. According to current core research, cancer patients must receive precise and effective treatment if the malignant locations and cancer class are correctly determined.

Despite CNN's capacity to reliably recognize essential feature space in input samples, it cannot associate with the relationship between vital components and demands excess features. The capsule network, as an optimal alternator, solves most of the shortcomings of CNN. Using pose vectors (singular refinements), capsule networks parameterize the connections among component portions of capsules (predictors) and then link all of those portions to elevated objectives of concern in a hierarchical fashion. A capsule network-inspired deep-learning methodology is integrated into the research for MGECD-based cancer class prediction. The main objective of the research work is the accurate prediction of cancer disease, which aids in the reduction of mortality rates. A two-level data augmentation strategy called smart data augmentation is a major contribution to this research. In this strategy, UDA integrates with W-GAN and is applied to generate the synthetic samples. For selecting the most significant features in the gene data, the correlation-based feature selection technique is utilized. The CapsNet deep learning model is used to train and classify the gene data and to predict a better outcome.

This research work is divided into six sections, the first of which contains the essential introduction to the proposed model. Section II describes the relevant studies that focus on cancer disease prediction. Section III delineates the dataset utilization and procedures of the proposed model and discusses feature selection. Section IV briefly delineates the core concept and utilization of CapsNet. Section V discusses the results and performance of the proposed model concerning data augmentation, accuracy, recall, and precision. It also signifies the impact of CapsNet in the prediction of cancer disease classification. Finally, Section VI concludes the research with a short summary of the overall performance of the proposed model.

II. RELATED WORK

The extraction, classification, and prediction of target-class features have reached higher levels because of developments in deep learning. DL networks differ from other ML algorithms in terms of prognosis. Deep networks have the ability to handle large volumes of data. But the most time-consuming parts of DL networks are training and retraining, which require high-performance platforms. The scientific community has taken a number of initiatives, particularly in the field of cancer prediction. This section covers the literature related to cancer disease prediction.

The authors [6] proposed that nuclei recognition from tumor histopathological imaging was made easier using a DL technique known as the stacked minimal auto-encoder (SMAE). As it accumulates high-level differentiating traits from sample intensity alone, SMAE is able to solve the issue of the nuclei's variety in shape, morphology, appearance, and texture. In order to characterize every patch of an image as either non-nuclear or nuclear, high-dimensional characteristics are extracted from those images and provided to the predictor. The SMAE nuclei identification method did better than other standard methods. It got an improved F-measure of 84.49 percent and an average precision-recall curve of 78.83 percent.

The authors [7] used an adaptive filtering covariance classifier to categorize fourteen categories of biological cancer cell type images, seven of which were associated with breast cancer and the other seven with liver cancer. The scientists used dual-tree multi-wavelet transform estimates since margins seem to be the main significant aspect in the visuals of the cancer cell strains studied. It is possible to define the margins in numerous orientations using this methodology, which enables the detailed study of cancerous cell line structure from image datasets. Since the images in the collection include a considerable number of pixel intensities, segmentation is conducted prior to feature collection. The EM (Expectation-Maximization) method is used only after the image has been modeled as a combination of two non-linear Gaussian functions and the consequent noisy patterns have been created. The noise is reduced by median filtering (MF) techniques and morphological procedures involving ''closure.'' The images are randomly divided into square frames, and the correlation matrices for each window frame are calculated. With a 98% success rate, the generated vectors were used to train an SVM ML model with the required kernel function on a set of nearly 840 images.

Binarization is a conventional method that turns a digital image into a binary image in order to minimize and convert the image dimensions. The main challenge in binarization is limited efficiency and accuracy degradation. There are several approaches to overcome the limitations of binarization techniques: BinaryDM, BiMatting, and BiBench. The binary diffusion model (BinaryDM) is used to earn the parameters by binarization and improve the optimization direction with low rank representation [8]. BiMatting is a novel binarization approach for accurate and efficient video matting.

In the encoder block, the shrinkable biniraized technique is utilized for extracting the most meaningful information. In the decoder block, sparse feature selection is selected [9]. To deal with minimizing the bit width and saving memory, the novel technique is the BiBench binarization technique. In this technique to deploy the ability and performance of biniraized networks, the binary operator plays a critical role. The accuracy value of binarization varies based on the different neural network architectures and learning parameters. It also requires limited hardware support [10].

A deep convolution neural network (D-CNN) was constructed and trained on 12,000,000 images (with high definition) by the authors [11] to categorize the cancer forms into 1000 distinct categories. This system comprises a total of five-convoluted layers that are also preceded by limit layers (max pooling) and three-completely linked layers. In the convolutional layers, a regularization approach called ''dropout'' is used to address the overfitting issue. The proposed methodology yielded top-1 and top-5 erroneous margins of 37.5 percent and 17 percent, respectively. So, in terms of overall performance, the result is much better than the relative methods that came before it.

To categorize cancer from histopathology datasets, the authors [12] suggested a method that incorporates medically comprehensible morphological characteristics, including 115 feature vectors such as textural and gray-level features, coloration, gray-level contour factors, and color-based components, wavelet, and tamuras parts. Before feature retrieval, an intensity-restricted dynamic histogram equalization approach is used to optimize the picture concerning intensity and stained pattern. The ROI (region of interest) is subsequently extracted from the dataset using the K-means classifier. Fuzzy K-Nearest Neighbor, Random Forest, and SVM classifiers are used to evaluate the method's effectiveness. According to the experimental investigations, K-NN surpassed all alternative classifiers in terms of effectiveness, susceptibility, and sensitivity.

The authors [13] proposed a diverse sample generation approach to enhance the diversity of the generated samples. The existing data-free quantization approaches, like quantization-aware training and post-training quantization, are used with and without training the original dataset. It consumes more time and increases the model's complexity. The suggested approach can be applied to various neural network architectures, particularly in situations with extremely low bit widths. The proposed scheme utilized three techniques: 1) Alignment of data distribution using batch normalization; 2) Layer-wise enhancement of the sample based on loss value; and 3) Diversification of the synthetic data using correlation inhibition.

To deal with the wide variety of histological pictures, the authors [14] suggested a novel classifier model that categorizes breast cancer tissue sample images independently of their resolution. In this setting, two distinct network designs have been developed: 1) a CNN that is optimized for a single task (specific), and 2) a CNN that is optimized for several (multiple) tasks. The cancer portion of the tissues from the data image can be predicted using the specific task CNN architecture; however, the multi-task architecture can simultaneously forecast both the carcinoma and the amplification factor of the available image. The most essential features of the approach are its flexibility for higher levels of amplification as well as its short training phase that covers all possible amplification factors. In addition, rotating and flipping training images is an eight-fold increase in batch size. The solitary classification algorithm obtained an overall recognition performance of 83.25 percent for cancer detection and amplification assessment. On the other hand, the average detection performance of the multi-task convolution layer was 82.13 percent for cancer prediction and 80.10 percent for amplification assessments.

On a publicly accessible dataset called ''Breast histology (ICIAR 2018),'' the authors [15] tested the capacity of a D-CNN to classify breast cancer histopathology images. The data were subjected to binary and multi-class categorization. Stochastic lengthening is used in the pre-processing stage to get the most out of the genes. Twelve non-overlapping segments (512×512) in all actual images are used to train standard CNN and hybridized CNN (CNN+SVM using the RBF kernel) classifications with randomized initial weightage scores, and patch-wise probabilities are calculated. Each image portion is rotated and mirrored to enhance the valid information. Using patch likelihood merger techniques such as optimum probability, the majority rule, and the summation of likelihood, it is possible to classify images based only on their contents. In this study, the best and worst methods for combining probabilities were determined to be the majority rule and the optimum probability technique. For multiple classifications, CNN plus SVM outperformed the traditional CNN approach with an efficiency of 77% and also reported an efficiency of 83.3% for binary categorization.

The authors [16] employed a pre-computed network model known as ''CaffeNet,'' primarily focusing on feature selection and extraction, in order to address the issue of rising intricacy and intensive training durations. They used linear interpolation as a predictor. The purpose of constructing a cognitively effective solution is the motivation for reusing the design and parameters of a pre-computed CaffeNet classifier. It was shown that using a pre-computed network as a feature representer is a good alternative to manual feature labels and some event-driven CNNs when making a very accurate cancer detection algorithm.

Using a feature pack technique, the authors [17] analyzed a collection of 1502 cancer-related histopathology images and classified them into 18 different groups. Feature descriptors like ''raw block'' are used in this scenario for the purpose of locating important components within the data images that make up the training batch. A graphical dictionary or coding scheme comprising 150 characteristics was produced using the K-means classification method. The authors also utilized

TABLE 1. Neural network based cancer prediction models.

Reference	Network Type	Predictor	Layers Utilized	Features	Classes	Metrics
[22]	Stacked Denoising Autoencoder	Neural Network	Two	70	Two	Accuracy
[23]	Variational Autoencoder	Statistical Analysis	Two	100	Two	P-Values
[24]	Stacked Denoising Autoencoder	SVM and NN Classifiers	Two	500	Two	Sensitivity, Specificity Accuracy, Precision,F1-Score
[25]	Sparse Autoencoder	SoftMax regression classifier	One	1047	Two	Accuracy
[26]	Stacked Autoencoder	Characteristic Classification	One	100	Seven	Accuracy
[27]	Stacked Autoencoder	AdaBoost	Four	64	Two	Sensitivity, Specificity Accuracy, MCC
[28]	Regularized Autoencoder	SVM+K-Means	Two	37	Four	C-index, Log-rank p-value, Brier Score
[29]	Variational Autoencoder	Logistic regression	Two	100	Five	Accuracy
[30]	Contractive Autoencoder	Statistical Analysis	One	22	-	Cox regression, Frieda man test, Holm Step-down

terms like ''frequency'' and ''variational text periodicity weighting factor'' to create a new synthetic image depiction. There are two kernel features in SVM: histogram convergence and RBF. Research shows that proposed codebooks use fewer code units than raw-block codebooks to represent visual patterns. The descriptors of raw-block had better F-measure performance than the standard descriptors, despite the fact that the raw-block signifiers required additional code units. Additionally, the RBF kernel was shown to be a good fit for maximum descriptors because it increases accuracy. This technique has a considerable benefit by adapting to a single image's content.

The authors [18] suggested different deep learning hybrid models to classify the eight types of cancer samples and one normal sample of gene expression data. The authors compared the results in various hybrid models like one-layer and two-layer CNN, LSTM-RNN, and GRU-RNN. For tuning the hyperparameters, they utilized the Bayesian optimization algorithm. The two-layer GRU-RNN model achieved the highest classification accuracy of 97.8 percent when compared to other hybrid models. To identify the most significant genes associated with cancer disease, the authors [19] suggested a fuzzy gene selection approach. This approach to selecting the important features is based upon three feature selection techniques: chi-square, F-classif, and mutual information. These techniques were combined to select the single best score for the particular gene. The novel fuzzy classifier to predict cancer disease is based on various deep learning and machine learning methods like SVM, LR, and MLP. According to the results, the classification results will be enhanced when a fuzzy approach is utilized.

Feature selection and extraction techniques were applied by the authors [20] before classifying the genome datasets. Leukemia, colon, and prostate gene expression datasets were utilized to predict the cancer disease. In feature extraction, principal component analysis was utilized to extract the relevant information from the high-dimensional data. In the feature selection technique, modified particle swarm optimization was utilized to select the most significant features from the gene data. Finally, machine learning models like SVM, KNN, and Naïve Bayes are utilized for training and classification tasks. The classification accuracy values are improved after feature selection and extraction techniques. The authors [21] proposed a novel approach for selecting the significant features from the gene data. This approach integrates the sine-cosine algorithm with the cuckoo search algorithm. The authors also utilized a minimum redundancy and maximum relevance filtering strategy for gathering the relevant feature set. The SVM, NB, and KNN classifiers were utilized in the classification task. The SVM can achieve better results when compared to other techniques. Table 1 describes some related work for neural network-based cancer prediction models.

III. METHODOLOGY

A. DATASET

We selected the Gene Expression dataset from [31] to test the performance of our proposed model because it contained a vast collection of features (>14124). The gene data was organized into six classes, which allowed us to evaluate how effectively the proposed approach performed on specific types of data. Researchers from all over the world usually utilize gene samples from the databases of significant bioinformatics laboratories that provide advanced learning to predict cancer types based on the MGECD dataset. Microarray data is commonly employed in cancer research, where rapid screening of cancer conditions is critical in determining the type of therapy and its prognosis. The gene dataset contains critical facets of various cancer diseases, such as lung, breast, brain, endometrial, and prostate cancer. A ''DNA microarray'' is a research tool used in the medical industry that may measure many genes' interpretations

simultaneously. The use of microarray profiling to diagnose and classify tumor development is now widely accepted. There are a few improved approaches to evaluating microarray data for cancer diagnosis among the multiple available methodologies. In order to assess the overall impact of the proposed approach, we compare the accuracy implications of various current datasets with the MGEC dataset. While there are several datasets that have been utilized in different studies, Table 2 contains those that are specific to a particular cancer type.

TABLE 2. List of datasets.

Reference	Number of features	Disease Prediction
[32]	12600	Prostate
[33]	22283	Breast
[34]	12533	Lung
[35]	9821	Brain
[36]	8034	Endometrial

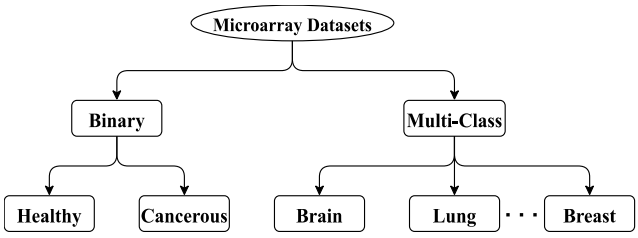

FIGURE 2. Types of microarray datasets.

The two different types of microarray datasets that are shown in Figure 2 are called binary and multi-class datasets, respectively. Binary datasets are typically utilized to discriminate between those who are malignant and those who are not. Alternatively, multi-class datasets, which can be difficult to work with and time-consuming, are utilized to distinguish between different types of cancer.

When displaying gene expression levels, MGECD is used to produce tables where rows represent samples (n), such as tumors, normal tissue, or test conditions, and columns represent genes (m). The values (Vpq) in each block (cell) represent the level of feature expression of a particular gene (Z) corresponding to a particular sample. The typical MGECD tabular format is shown in Figure 3.

B. SMART DATA AUGMENTATION

A trained model with few training sets has limited ability for generalization and is highly dependent on its original inputs. Extension of the training data has been widely used to avoid overfitting (inadequate generalization) concerns, which is commonly referred to as data augmentation. It also allows for the use of a larger connectivity system without causing excessive co-linearity. Data augmentation is typically employed when the samples from the source datasets are slightly altered and new samples are created artificially. Smart

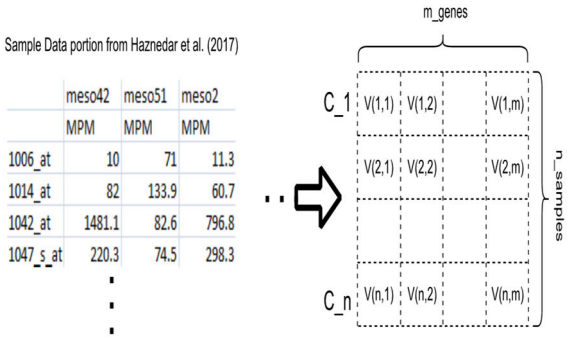

FIGURE 3. MGECD tabular format and its matrix form.

Augmentation (SA) is part of the research strategy. Smart augmentation involves two levels of enhancement to ensure the consistency and stability of newly generated data samples. The first level of the augmentation procedure is conducted based on uniform distributive augmentation. The second level encompasses the process of W-GAN, where augmented data samples are trained through a generator.

In contrast, the discriminator distinguishes the actual data from the generated data. W-GAN outperforms conventional oversampling techniques in terms of data quality enhancement. As contrasted to regular GAN, W-GAN calculates the difference between the distributions of the original data and the synthesized data using the Wasserstein distance [37]. Thus, the goal of this type of data improvement technique is to provide highly refined, appropriate data samples for the training set of the final classifier. Figure 4 represents the overall process sequence of the proposed model.

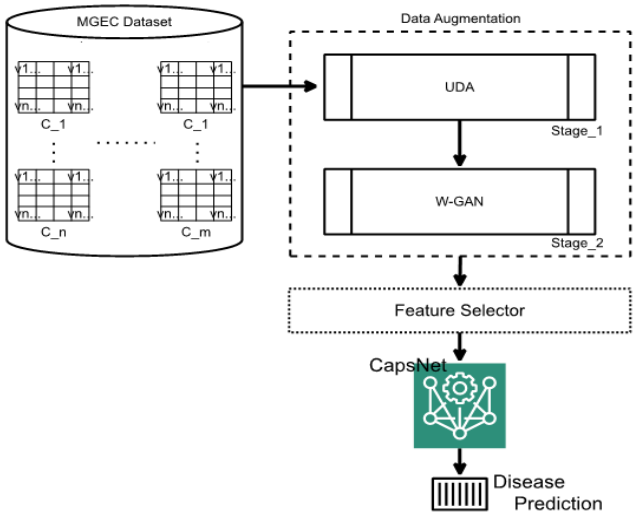

FIGURE 4. Proposed architecture of disease prediction.

1) FIRST-LEVEL AUGMENTATION: UNIFORM DISTRIBUTIVE AUGMENTATION (UDA)

We discuss our UDA techniques in this subsection as a first-level data augmentation mechanism. The UDA process is shown from a high-level perspective in Figure 5, which

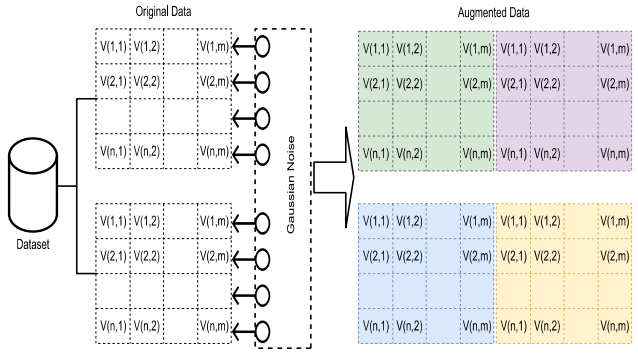

FIGURE 5. UDA strategy.

highlights the workflow. The underlying principle of our method is simple: we synthesize a newer sample based on the availability of class labels and their associated samples. Initially, the lowest recorded samples of each class label (C_i) are chosen for synthesizing purposes to match up with the count of the highest available samples of other C_i in the dataset. Thus, the process normalizes the sample count of each class label in the entire dataset through uniform distributive augmentation. The process of data augmentation involves the creation of duplicate rows of data with just a small amount of Gaussian noise applied. Here, the Gaussian noise is generated with the same dimension as a particular sample. The probability density function of Gaussian noise follows the Gaussian function. An improved prediction can be obtained by averaging the predictions from the copied rows. In Algorithm 1, it represents the procedure of noise inclusion into the MGECD tabular rows.

Algorithm 1 Noisy Procedure

Input: Scalar Vector (δ), Sample (Si)
Output: Noisy Data
1:Set δ;
2: Compute ß(standard deviation) of Vpq;
3: ∀ (1) Do //Instances
4: ∀ [(Vpq)|Z] Do
5: Choose random number (R), where R∈ (-ß, ß)
6: Incorporate R/δ →1;
7: End ∀
8: End ∀

An analytical term for the kind of added noise is always referred to as the Gaussian distribution, which has a probabilistic model that is identical across all points in the standard normal distribution sampling. Because it is so essential and prevalent in analytics, normal distributiveness is an appropriate first step. Although the noise generated here seems to be random, there is a pattern to it, as with any random occurrence. Because random events with a Gaussian distribution are still so prevalent in the empirical condition, the DL and data analytics fields strongly prefer the Gaussian distribution. A Gaussian variable seems to be any randomized parameter that would sum up several distinct random occurrences. Gaussian noise may be introduced to data collection in the following way:

Step 1: The variable is assigned with a randomized noise by generating a random amount of noise.

Step 2: The created noise should be added to the MGECD rows.

Step 3: The probability distribution of a Gaussian function can be determined using the given equation.

$$f_F(r) = \left[1 \Big/ \sqrt{2 \times \pi \times \sigma^2}\right] e \left[-0.5\left(\frac{(r-m)^2}{\sigma^2}\right)\right] \tag{1}$$

An arbitrary parameter F seems to be uniformly distributed when its probability distribution has a mean and variance, as stated in equation (1). As long as there is minimal potential for conflict between the arbitrary parameter F and its actual argument r, f(r) is merely used to signify the probability distribution's actual argument, r. In statistical practice, the Gaussian probability distribution of F is represented as follows:

$$F \sim \mathcal{N}\left(r - \left(\pm m, \pm \sigma^2\right)\right) \tag{2}$$

2) SECOND-LEVEL AUGMENTATION: WASSERSTEIN GENERATIVE ADVERSARIAL NETWORK(WGAN)

The W-GAN is used to construct the synthetic fake data samples [38]. W-GAN is based on the basic GAN concept. Therefore, it is essential to understand the basic working mechanism of the GAN. With GANs, dual adversary networking techniques (g(f), d(f)) are employed. While one network generator g(f) trains to create a synthetic sample based on the discriminator's d(f) feedback, d(f) distinguishes between the synthetic and real samples. The generator's purpose is to optimize a cost value Ψ (g(f), d(f)), but the discriminator's objective is to elevate it [39], [40].

Even though this approach may produce very good outcomes, it is still unknown how to objectively evaluate generative models. Techniques that produce high-quality sample data could have a low probability (the possibility that the learned information can be used in such a system). Conversely, methods with a high probability might produce inadequate data samples.

Thus, our major augmentation technique is the W-GAN, which is a development of the GAN and offers more consistency. W-GAN uses Wasserstein distance to directly calculate the distribution of the original and generated data, which is further, specified by infimum (the maximum lower limit) and stated as,

$$W\left[\mu_i, \mu_j\right] = \underset{\rho \in \varphi(\mu_i, \mu_j)}{Inf} \left([\|a - b\|] \cdot E_{(a,b)\sim\rho}\right) \tag{3}$$

where φ (μ_i, μ_j) indicates a possible joint distribution of ρ (a, b) whose extreme parameters are original distribution μ_i and generated distribution μ_j, respectively. 'φ' refers to all possible ρ distributions.

Furthermore, W-GAN provides a gradient descent (loss function) that is directly correlated with the accuracy and reliability of the samples that are created. It's one of the best and most effective ways to prevent GAN loss. As a result, mode collapse and the vanishing gradient are successfully handled. An alternate name for the discriminator in W-GAN is critic because the generated samples have been evaluated instead of precisely classified as true or false. The critic loss function is expressed in equation (4) and is employed to train the critic network, whereas the generator loss function is expressed in equation (5).

$$C_L = \left[\bar{C}_s(i) - \bar{C}_s(j)\right] \tag{4}$$

$$\mathcal{G}_L = -\left[\bar{C}_s(j)\right] \tag{5}$$

The critic network's output layer's sigmoid activation function is replaced with a linear one. The outcome does not have to be in the range of 0 to 1, due to this simple modification that inspires the critic to provide a score rather than a probability related to the data distribution. The critic output layer is activated in a linear fashion rather than a sigmoid one. As a result of such minor modifications, the critic prefers to produce a score instead of an associated possibility correlated to the generated data distribution. It also ensures that the output cannot fall anywhere between 0 and 1. Since the outcomes of the generator and the critic are not in accordance with the probabilistic measures (from 0 to 1), it is crucial to boost the actual (absolute) divergence between the outcomes of both networks, especially during the training process of the critique network. Likewise, the generator function's actual value is maximized throughout the retraining process of the generator network. Figure 6 represents the working mechanism of W-GAN.

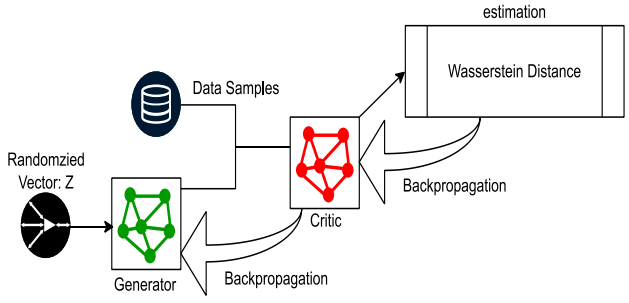

FIGURE 6. Working mechanism of W-GAN.

a: GENERATOR

After multiple training cycles, the generator starts generating synthetic data samples once the convergence of the loss functions of the generator and critique is achieved. The generator network includes three layers, which are depicted in Figure 7. By default, two neurons serve as the z vector's input, followed by three hidden layers composed of 512 neurons each. Finally, two neurons serve as the system's output. ReLU is the activator for the three hidden layers, while linear is the activator for the output layer.

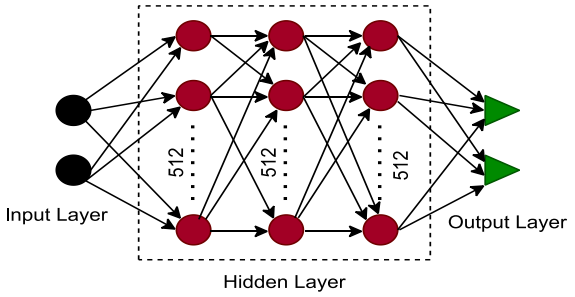

FIGURE 7. W-GAN generator process.

b: DISCRIMINATOR

Using the GAN's [41] approach, the critic's goal was to identify differences between the real samples and the fake ones, while the generator tried to produce artificial data samples that might confuse the critic. In other words, an artificial sample misleads the critic into assuming it to be real. The preliminary artificial samples that were given to the critic network were the source of the finalized synthetic data samples. Like a generator network, the critique network comprises two neurons that serve as the input layer, followed by three hidden layers composed of 512 neurons each. Finally, the output of the system is one neuron. The output layer uses the linear activator, while the three hidden layers utilize the ReLU activator. The usage of RMSprop optimizer helps reduce the instability risk that Adam optimizer often presents during training. Figure 8 shows the process of critic network.

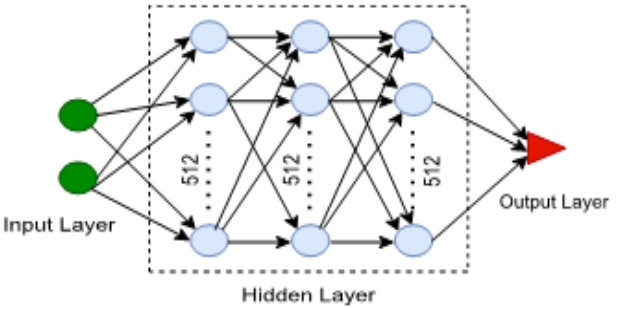

FIGURE 8. W-GAN discriminator process.

Consequently, this ensures that the gradient won't vanish. Applying the loss function encourages the generated samples to converge toward the originals.

C. FEATURE SELECTION AND FILTER

The correlation technique is used in this research to choose relevant features. A correlation-based computation process is used in this approach to compute the correlation coefficients of various gene features. It works over the limitations of univariate filter methods, which ignore how different gene components (features) interact. For instance, correlation factors can be utilized for investigation into the relationship between different gene expression pairings. Pearson's CC, which is often referred to as correlated indices in analytics,

is used to find a linear relationship between various gene pairs. The evaluation of the amount of significance or connection between the two variables is referred to as a ''correlation''. The CC of both traits in a linear relationship is ±1. The CC is 0 if there is no correlation between the characteristics. To find the linear CC (R) for a particular set of gene variables, use the following expression (6) from [42].

$$\boldsymbol{R} = \sum \left[\left(\boldsymbol{V}_i - \bar{\boldsymbol{V}}_i \right) \left(\boldsymbol{U}_i - \bar{\boldsymbol{U}}_i \right) \right] \Big/ \sqrt{\sum \left[\left(\boldsymbol{V}_i - \bar{\boldsymbol{V}}_i \right)^{\mathbf{2}} . \left(\boldsymbol{U}_i - \bar{\boldsymbol{U}}_i \right)^{\mathbf{2}} \right]} \tag{6}$$

Using Pearson's correlation coefficient (PCC), significant features are determined. The linear relationship between two variables, U and V, can be estimated numerically. The boundaries executed by R on the variables are conditionally represented as $[-1 \leq R \leq 1]$. In other words, a positive correlation is indicated when R approaches 1, a negative correlation is indicated when R approaches a negative value, and a non-linear relationship is indicated when R approaches 0. Pearson's correlation was used to calculate the degree of relationship between the various relevant features in the proposed model. Typically, features are selected if the score is above the cutoff value ($\zeta = 0.5$). When the feature count equals (n-log n), the selection is complete. Finally, the feature-to-feature interaction technique is used to remove irrelevant and extraneous features.

The removal of redundant features is the primary intent of this method. Initially, several fundamental concepts concerning selection processes are referred to from [43]. The total feature set is indicated as |F|, and the gene feature is denoted as f_i. Consequently, $r_i = [|F| - \{fi\}]$ may be used to calculate the relevance r_i, and the conditional likelihood (L) of cancer class (k) for a given f_i can be written as,

$$\boldsymbol{f}_i = \begin{cases} \boldsymbol{relevant}, & \exists \boldsymbol{r}_i' \subseteq \boldsymbol{r}_i, \, (\boldsymbol{i.e}) \, \boldsymbol{L} \left[\left\langle \boldsymbol{k} \mid \boldsymbol{f}_i, \boldsymbol{r}_i' \right\rangle \right] \\ \quad \neq \boldsymbol{L} \left[\left\langle \boldsymbol{k} \mid \boldsymbol{r}_i' \right\rangle \right] \\ \boldsymbol{irrelevant}, & \boldsymbol{Otherwise} \end{cases} \tag{7}$$

Equation (7) states that it is only significant if taking a f_i out of a |F| reduces the predictive performance. The statement indicates that a f_i can be determined to be relevant based on two characteristics,

- It is strongly related to the desired target class.
- It forms a subset with other f_i, and this f_i-subset is strongly related with the goal notion.

Thus, feature interaction happens if f_i is significant as a result of the second argument.

IV. CAPSULE NEURAL NETWORK

In this research, Capsule Neural Network (CapsNet) is incorporated for classification as well as prediction purposes. CapsNet are one of the advanced forms of DL wherein the core concept is incorporated in structuring and classifying disease. CapsNet apply a dynamic routing technique to pinpoint the direction for information that is needed to be sent. Based on the current studies, it seems that the network's functionality is meeting several desired expectations. But it's still not apparent how well the CapsNet operate and perform in computer-aided diagnosis. Dynamic routing processes are incredibly intensive, keeping these networks much more delayer in processing than other recent DL networks. Multi-dimensional features, especially from MGECD, complicate the utilization of CapsNet. In addition, it is unknown regarding the base operations of CapsNet to compare against other cutting-edge technologies, especially in the medical sector. Besides these, the vital contributions are made in response to any investigation relevant to the medical field; some are,

- A variety of cancer classification tasks are evaluated using CapsNet.
- Whenever the CapsNet perform with massive data samples, they outperform CNNs, which is a notable fact.

A. COMPUTATION PROCESS

In the conventional design of CapsNet, the term ''capsule'' refers to a collection of neurons (an activating vector is a set of outcomes that are linked together). They provide predictions about certain objects' existence and the posture characteristics associated with a specific gene data point. A dynamic array can capture the perspective position of an object, and the activation matrix of each vector length (standard or significance) can be used to approximate the existence of the particular object. For example, the length and strength of the activation matrices of various vectors do not vary when a sample is rotated. The length of the outcomes from lower capsule/units (l_1, l_2,..., l_n) convey the likelihood that a specific entity exists. In addition to encoding vital qualities like dimension, direction, and current status (position), the vector directions might convey other attributes too.

A linear mapping function (transformation matrix: T_{ij}) is computed to represent the connection between the i^{th} capsule of the bottom layer and the j^{th} capsule of the upper layers. Thus, the data point is propagated as $d\,\hat{}_{(i|j)} = T_{ij} d_{ij}$. Here, $d\,\hat{}_{ij}$ denotes any i^{th} capsule's belief in a bottom layer over the j^{th} capsule of upper layers. For instance, $d_{j|1}$ reflects the expected gene expression needed to classify the target classes. The CapsNet will eventually learn enough about the transformation matrix of each capsule pairing to store (encode) the aspectual relationship between them.

B. DYNAMICROUTING

Once the predictive vectors have been constructed, the capsules of the bottom layer transmit their data to the parental neurons that most closely match their forecasts. When a child capsule's outcomes are routed to their respective parental capsules, it is known as Dynamic Routing. The transit coefficient R_{ij} between the i^{th} bottom layer capsules to the j^{th} upper layer capsule is denoted as $\Sigma_j R_{ij} = 1$, and here $\forall_j$, $R_{ij} \geq 1$. There will be no communication between the ith capsule and the j^{th} capsule, if $R_{ij} = 0$; but there will be communication if $R_{ij} = 1$. Child capsules' responses are sent sequentially to the suitable

succeeding capsule, thus ensuring the acquisition of correct data resulting in a much more precise estimation of the gene expression patterns. Bottom layer capsule deciding whether or not to convey its information to the parental capsules. Predictive vectors are scaled by an appropriate value of the R_{ij} to arrive at a final conclusion. The following equations (8) and (9) delineate the computation process of the parental capsules ($\mathbb{P}$j) as well as R_{ij} in CapsNet [44].

$$\mathbb{P}_j = \left[\frac{s_j}{\|s_j\|}\right] \cdot \left[\frac{\|s_j\|^2}{1+\|s_j\|^2}\right], s_j = \sum_i \hat{d}_{i|j} \cdot R_{ij} \quad (8)$$

$$R_{ij} = exp\left[b_{ij}\right] \Big/ \sum_k exp\left[b_{ij}\right], b_{ij} \rightarrow b_{ij} + \hat{d}_{\langle j|i\rangle} \cdot \mathbb{P}_j \quad (9)$$

Squash (s) non-linearity is applied to the outcome of every parental capsule $\mathbb{P}$j, which is the weighted summation of total estimates from the bottom layer capsules. There is no change in the while squashing; therefore, the length of the result vector may be regarded as the chance of a specific feature being recognized by each capsule (usually >1). Parental capsules get forecasts from all child capsules. Such vectors are shown as data points. It will boost the jth parental capsule's R_{ij} by a magnitude of d ^$_{\langle j|i\rangle}$· P_j. Child capsules convey additional data to their $\mathbb{P}_j$ with outputs that are closer to their predictions d ^$_{\langle j|i\rangle}$ than those with less comparable outcomes.

V. RESULTS AND DISCUSSION

The proposed model's performance has been evaluated in comparison with a few existing models like Fuzzy-KNN, CaffeNet, and CNN+SVM, which are already discussed in Section II. The proposed approach incorporates a two-stage smart augmentation process for data augmentation, which is the first major evaluation of this research. To increase the data samples in MGECD is represented as 'n' which affects the accuracy of predictions.

TABLE 3. Evaluation of data augmentation at different stages of smart augmentation.

MGECD	Original Size	UDA Strategy	W-GAN	Smart Data Augmentation
Lung	189	232	235	238
Prostate	205	241	239	240
Endometrial	169	216	220	227
Breast	247	278	267	271
Brain	206	245	248	250
Healthy	190	257	253	256

With an increasing size of n, the synthetic data sample diversifies. If 'n' is too large (significant deviation), it is possible to obtain samples that are not representative of the original data. Conversely, the generated data lacks variance if n is kept at a low level. To ensure that CapsNet has access to sufficient samples for training and analysis, the total sampling size has been increased to 1/3rd of the overall sample. The sample size generated at 2-stages (UDA and W-GAN) is compared with the sample size generated after the execution of the 2-stage, which is referred to as smart augmentation. Table 3 and Figure 9 show the variety of distinct samples produced by the smart augmentation strategy.

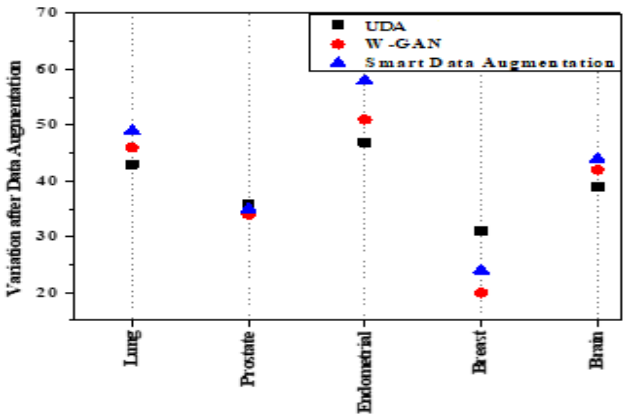

FIGURE 9. Augmentation variation at different level of smart data augmentation process.

The fake samples generated using the UDA strategy and W-GAN against samples generated after the entire augmentation process (smart segmentation) are shown in Table 3. Compared to the 2-stage augmentation, the samples are further enhanced after the completion of smart augmentation with a variation of 4.84% against the UDA strategy and 7.54% against the W-GAN, respectively. The scatter plot from Figure 9 exhibits the variation of data sample augmentation for five different categories and the variations in the generated data samples between different augmentation techniques. This result states that there is a maximum variation between W-GAN and smart augmentation because the samples have to undergo the augmentation process as a repeated one but still produce maximum samples than the noise mixing strategy for all five classes.

In cancer disease prediction, the impact of data increases on recall values is investigated. In order to increase the sample data size, a UDA strategy and W-GAN methodologies were combined used. Figure 10 exhibits that the recall value for the two-staged augmentation technique diminishes beyond 40%. When studied and observed separately, W-GAN and noise added sample via UDA create duplicated and redundant data. Concentrating data point on a single dataset reduces diversity in the dataset. The combined outcome of both processes as smart augmentation generates distinctive synthesized samples with minimum deterioration compared to the individual performance of the other two stages. In terms of the probability distribution, the synthesized samples are quite excellent. As a result of using a smart augmentation strategy results in a greater proportion of distinct synthesized samples. The accuracy, precision, and recall values of the proposed method are evaluated in the following equations (10), (11) and (12).

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (10)$$

$$Precision = (TP)/(TP + FP) \quad (11)$$

$$Recall = (TP)/(TP + FN) \quad (12)$$

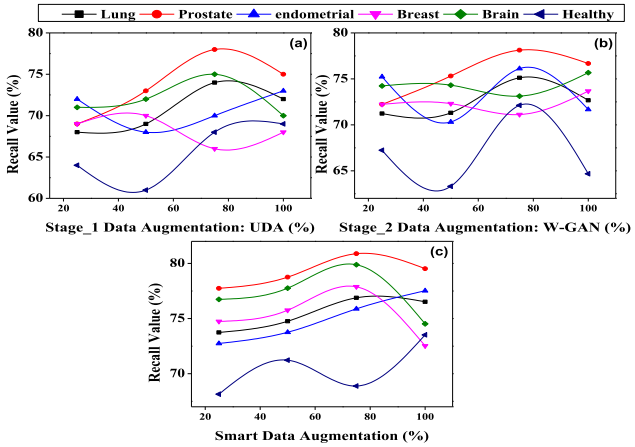

FIGURE 10. Recall value vs augmentation percentage.

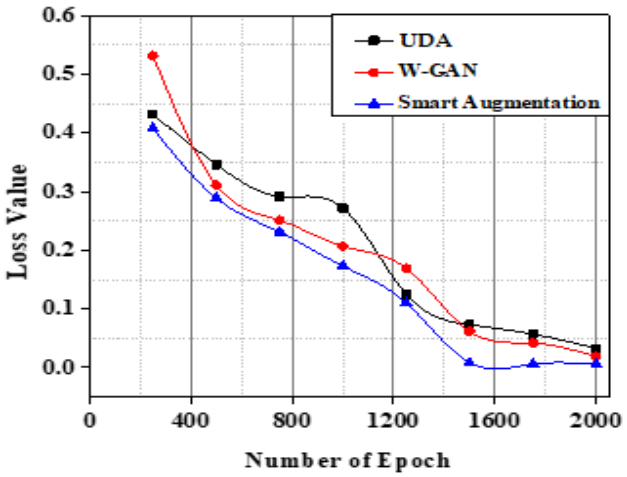

FIGURE 11. Graphical illustration of loss curve.

TABLE 4. Predictive results when using UDA.

Methods	Precision	Accuracy	Recall	AUC
Fuzzy-KNN	89.1	90.35	88.17	0.87
CapsNet	97.14	98.03	98.47	0.95
CNN+SVM	92.13	92.12	91.03	0.89
CaffeNet	93.25	94.56	93.45	0.91

TABLE 5. Predictive results when using W-GAN.

Methods	Precision	Accuracy	Recall	AUC
Fuzzy-KNN	96.32	95.42	91.1	0.88
CapsNet	98.29	98.17	97.58	0.98
CNN+SVM	96.15	96.11	93.24	0.91
CaffeNet	97.23	95.22	94.21	0.93

TABLE 6. Predictive results when using smart data augmentation.

Methods	Precision	Accuracy	Recall	AUC
Fuzzy-KNN	97.34	96.44	92.12	0.91
CapsNet	100	99.32	98.56	0.98
CNN+SVM	96.53	95.41	94.33	0.96
CaffeNet	97.09	96.37	95.16	0.97

Table 4, 5, and 6 compares the outcomes of efficient classifiers, such as Fuzzy-KNN, CapsNet, CNN+SVM, and CaffeNet. The results from the above tables also exposes the impact of three different data augmentation on four classifiers in terms of precision, accuracy, Recall, AUC (Area Under Curve). The accuracy of each classifier increases when using the proposed smart data augmentation compared with the UDA and W-GAN approaches separately. Though there are no major variations in the outcome among three different augmentation techniques, smart augmentation exceeds both UDA and W-GAN by showing mild variation of 2.28% in precision, 1.32% in accuracy, 0.54% in recall, and 1.5% in AUC.

Figure 11 depicts the evaluation of the loss function of three augmentation techniques. All the three augmentation techniques are iterated for 2000 epochs and the value for the loss function reduces from 0.432 to 0.033 for UDA strategy, from 0.532 to 0.019 for W-GAN, and from 0.408 to 0.005 for smart augmentation as shown in the Figure 11. It is clear from the results that W-GAN performs better than the UDA technique. This is because the generator learns from the critic network at every stage, and the feedback it receives from the critic helps it produce samples that are extremely similar to the original data. But when both the technique is employed together (smart augmentation), the outcome is further refined and enhanced. This guarantees that the produced data has a high sensitivity value, making it suitable for use in crucial fields like healthcare.

The AUC (Area under Curve) is a crucial tool for determining and evaluating the constructed classifier's accuracy. A model's significance and accuracy are enhanced by executing this assessment. This tends to boost the detection rate. Predictive models with variable probability limits sometimes face a tradeoff between the precision of their classification performance and the accuracy of their diagnostic relevance.

The classifier accuracy of the three classifiers, namely, Fuzzy-KNN, CNN+SVM, CaffeNet are compared with the CapsNet for in-depth analysis on the prediction of cancer types. According to Figure 12, CapsNet performs better than the accuracy value with the minimum false positives among the classifiers listed. The findings suggest the accuracy value might be much improved. It is also proven that both the specificity and sensitivity predict the prevalence of disease via available trained datasets. Still, CapsNet with smart augmentation produces high sensitivity (test's capacity to accurately identify the individuals having a cancer disease). Similarly, the outcomes also exhibit the high specificity for CapsNet (test's capability to accurately identify healthy individuals).

As depicted in figure 13, the prediction of each class is expressed by a diagonal column in the confusion matrix, and all other components of the matrices are zero, indicating that no samples were misclassified. Some off-diagonal positions

TABLE 7. Predictive analysis of cancer subtypes using different datasets.

Cancer	Sub Types	ALL	Brest	Prostate	Lung	Brain	Endometrial
Breast	Luminal A or HR+/HER2- (HR-positive/HER2-negative)	0.97±0.03	0.91±0.05	NA	NA	NA	NA
	Luminal B or HR+/HER2+ (HR-positive/HER2-positive)	0.98±0.02	0.89±0.08	NA	NA	NA	NA
	Triple-negativeor HR-/HER2-(HR/HER2-negative)	0.98±0.04	0.92±0.03	NA	NA	NA	NA
Prostate	TP53	0.97±0.01	NA	0.85±0.08	NA	NA	NA
	P13K	0.96±0.05	NA	0.91±0.03	NA	NA	NA
	ETS	0.98±0.04	NA	0.88±0.02	NA	NA	NA
Lung	Bronchoid	0.95±0.09	NA	NA	0.92±0.03	NA	NA
	Magnoid	0.97±0.02	NA	NA	0.87±0.05	NA	NA
	Squamoid	0.94±0.02	NA	NA	0.89±0.03	NA	NA
Brain	Proneural	0.98±0.01	NA	NA	NA	0.84±0.03	NA
	Neural	0.96±0.03	NA	NA	NA	0.92±0.01	NA
	Classical	0.96±0.02	NA	NA	NA	0.90±0.04	NA
	Mesenchymal	0.97±0.07	NA	NA	NA	0.82±0.08	NA
Endometrial	POLE Ultramutated	0.95±0.06	NA	NA	NA	NA	0.91±0.02
	Microsatellite Instability Hypermutated	0.93±0.09	NA	NA	NA	NA	0.89±0.04
	Copy Number Low	0.95±0.05	NA	NA	NA	NA	0.87±0.06
	Copy Number High	0.97±0.03	NA	NA	NA	NA	0.83±0.04

*HER-2: Human Epidermal Growth Factor Receptor 2,
*HR: Hormone Receptors,
*ETS: Erythroblast-26 Transformation Specific,
*PI3K: Phosphoinositide 3-Kinase,
*TP53: Tumor Protein,
*POLE: Polymerase Epsilon, NA: Not Applicable

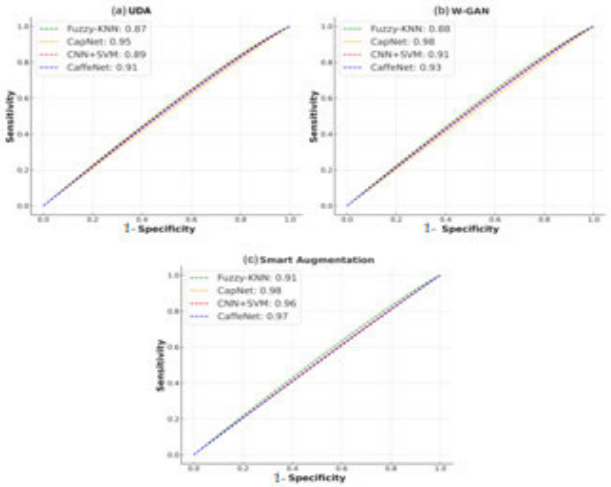

FIGURE 12. Performance evaluation via AUC.

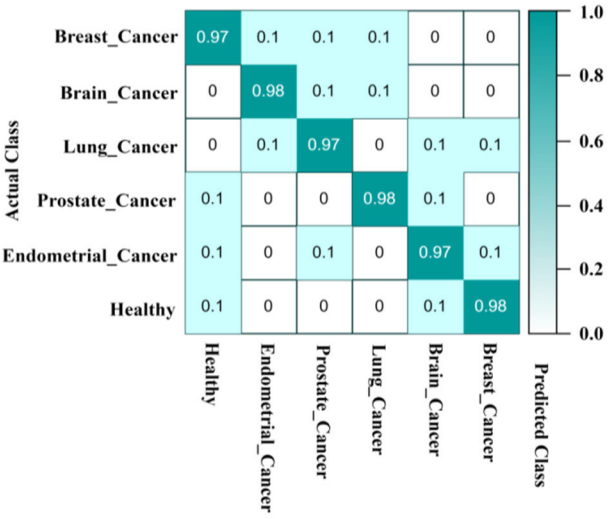

FIGURE 13. Performance of CapsNet using smart augmentation.

also include the incorrectly categorized class labels. The results show that the suggested model is ∼98 percent accurate in predicting all the required classes, which is impressive.

Table 7 states that the accounted augmented data samples from MGECD appropriately classified and predicted the subtypes of various cancer diseases. From the

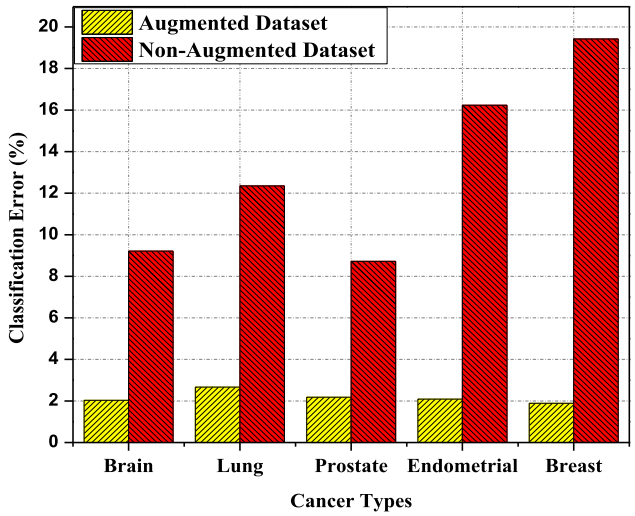

FIGURE 14. Augmented vs non-augmented dataset.

comparative analysis of the outcome obtained, it is noted that the accuracy difference for the subtype of breast cancer (Luminal A, Luminal B, and Triple-negative) is 7.40% compared to the existing dataset [33]. Similarly, the comparative results between prostate cancer subtypes (TP53, P13K, and ETS) from MGECD [32] expose an accuracy difference of 6.45%. Furthermore, the estimated outcome of the Lung cancer subtypes (Bronchoid, Magnoid, and Squamoid) exhibits a 9.87% accuracy difference compared with an existing dataset [34]. Likewise, the observed accuracy difference in predicting brain cancer subtypes (Proneural, Neural, Classical, and Mesenchymal) between MGECD and [35] datasets is 6.31%. Finally, the comparative analysis of predictive results of endometrial cancer subtypes (POLE Ultramutated, Microsatellite Instability Hypermutated, Copy Number Low, and Copy Number High) recorded an accuracy difference of 0.43% among [36] datasets. In all the predictions, the considered dataset in the proposed model excelled the existing datasets due to the inclusion of a 2-staged augmentation process.

Besides the experimental outcomes, it is essential to investigate the resultants of augmented and non-augmented datasets. Figure 14 exposes the comparative classification error report of both augmented and non-augmented datasets in predicting different cancer types. The outcome insists on the necessity of the augmentation process for the sparse datasets since the augmented dataset produces better results than the non-augmented dataset.

VI. CONCLUSION

Studies have shown that when it comes to making vital decisions, researchers require a large amount of information. Healthcare applications require a vast volume of data in order to get reliable results. MGECD was shown to be a data imbalance when there were fewer samples with a huge number of features. The researchers must utilize advanced methodology to create artificial data at this point since gathering more data samples under experimental conditions requires effort. Artificially generated data must be closely related to the original data in order to be seriously examined. The potential benefit of this research work is to enhance the sample size to solve the imbalance problem in the gene data and to make an appropriate class diagnosis of cancer with high accuracy using DL methods. Data is generated using a two-stage augmentation approach, which is referred to as "smart augmentation". The augmentation process comprises UDA strategy as the first stage and W-GAN as the second stage.

The results show that the loss function with smart augmentation has a significantly lower value. The generated data is assured to be sensitive. Recall, accuracy, and precision are employed to evaluate the efficiency of the classifier. The CapsNet classifier with a smart data augmentation strategy achieved the highest accuracy of 99.32 percent, recall of 98.56 percent, and precision of 100 percent when compared to other approaches examined. The results show that the performance improvement of the generator will lead to a significant increase in prediction accuracy. Thus, smart data augmentation could be effective for generating data for important applications. For the classifier to perform correctly, it must be fed accurate information. Larger sample sizes enable researchers to narrow their emphasis on identifying the optimum feature combination.

REFERENCES

[1] (2020). *Global Cancer Observatory*. Accessed: Dec. 4, 2023. [Online]. Available: http://gco.iarc.fr/today

[2] H. Sung, J. Ferlay, R. L. Siegel, M. S. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Medicine*, vol. 13, no. 1, pp. 1–17, 2021.

[4] U. Ravindran and C. Gunavathi, "A survey on gene expression data analysis using deep learning methods for cancer diagnosis," *Prog. Biophys. Mol. Biol.*, vol. 177, pp. 1–13, Aug. 2023.

[5] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inform.*, vol. 2, Jan. 2006, Art. no. 117693510600200030.

[6] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 119–130, Jan. 2016.

[7] F. Keskin, A. Suhre, K. Kose, T. Ersahin, A. E. Cetin, and R. Cetin-Atalay, "Image classification of human carcinoma cells using complex wavelet-based covariance descriptors," *PloS One*, vol. 8, no. 1, 2013, Art. no. e52807.

[8] X. Zheng, H. Qin, X. Ma, M. Zhang, H. Hao, J. Wang, Z. Zhao, J. Guo, and X. Liu, "Towards accurate binarization of diffusion model," 2024, *arXiv:2404.05662*.

[9] H. Qin, L. Ke, X. Ma, M. Danelljan, Y. W. Tai, C. K. Tang, X. Liu, and F. Yu, "BiMatting: Efficient video matting via binarization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–20.

[10] H. Qin, M. Zhang, Y. Ding, A. Li, Z. Cai, Z. Liu, F. Yu, and X. Liu, "Bibench: Benchmarking and analyzing network binarization," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28351–28388.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[12] R. Kumar, R. Srivastava, and S. Srivastava, "Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features," *J. Med. Eng.*, vol. 2015, pp. 1–14, Aug. 2015.
[13] H. Qin et al., "Diverse sample generation: Pushing the limit of generative data-free quantization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11689–11706, 2023, doi: 10.1109/TPAMI.2023.3272925.
[14] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2440–2445.
[15] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *IEEE Access*, vol. 6, pp. 24680–24693, 2018.
[16] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, "Deep features for breast cancer histopathological image classification," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 1868–1873.
[17] J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using bag of features and kernel functions," in *Proc. 12th Conf. Artif. Intell. Med.*, 2009, pp. 126–135.
[18] S. Babichev, I. Liakh, and I. Kalinina, "Applying the deep learning techniques to solve classification tasks using gene expression data," *IEEE Access*, vol. 12, pp. 28437–28448, 2024, doi: 10.1109/ACCESS.2024.3368070.
[19] M. Khalsan, M. Mu, E. S. Al-Shamery, S. Ajit, L. R. Machado, and M. O. Agyeman, "A novel fuzzy classifier model for cancer classification using gene expression data," *IEEE Access*, vol. 11, pp. 115161–115178, 2023, doi: 10.1109/ACCESS.2023.3325381.
[20] A. Razzaque and A. Badholia, "PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification," *Meas. Sensors*, vol. 31, Aug. 2024, Art. no. 100945.
[21] A. Yaqoob, N. K. Verma, and R. M. Aziz, "Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm," *J. Med. Syst.*, vol. 48, no. 1, pp. 1–14, Jan. 2024.
[22] A. J. Titus, C. A. Bobak, and B. C. Christensen, "A new dimension of breast cancer epigenetics," in *Proc. 9th Int. Conf. Bioinf. Models, Methods Algorithms*, 2018, pp. 1–15.
[23] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," in *Proc. Biocomputing*, Jan. 2018, pp. 80–91.
[24] P. Danaee, R. Ghaeini, and D. A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," in *Proc. Pacific Symp. Biocomputing*, 2017, pp. 219–229.
[25] L. Dyrskjot, T. Thykjaer, M. Kruhoffer, J. L. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolft, and T. F. Orntoft, "Identifying distinct classes of bladder carcinoma using microarrays," *Nature Genetics*, vol. 33, no. 1, pp. 90–96, 2003.
[26] J. Tan, M. Ung, C. Cheng, and C. S. Greene, "Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders," in *Proc. Pacific Symp. Biocomputing Co-Chairs*, 2014, pp. 132–143.
[27] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer," *IEEE Access*, vol. 6, pp. 28936–28944, 2018.
[28] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep learning-based multi-omics integration robustly predicts survival in liver cancer," *Clinical Cancer Res.*, vol. 24, no. 6, pp. 1248–1259, 2018.
[29] J. Liu, X. Wang, Y. Cheng, and L. Zhang, "Tumor gene expression data classification via sample expansion-based deep learning," *Oncotarget*, vol. 8, no. 65, 2017, Art. no. 109646.
[30] L. Macias-Garcia, J. M. Luna-Romera, J. Garcia-Gutierrez, M. Martinez-Ballesteros, J. C. Riquelme-Santos, and R. Gonzalez-Campora, "A study of the suitability of autoencoders for preprocessing data in breast cancer experimentation," *J. Biomed. Inform.*, vol. 72, pp. 33–44, Sep. 2017.
[31] B. Haznedar, M. T. Arslan, and A. Kalinli, "Microarray gene expression cancer data," Mendeley Data, V4, 2017, doi: 10.17632/ynp2tst2hh.4.
[32] D. W. Zimmerman and B. D. Zumbo, "Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks," *J. Experim. Educ.*, vol. 62, no. 1, pp. 75–86, Jul. 1993.
[33] A. S. Assiri, A. G. Hussien, and M. Amin, "Ant lion optimization: Variants, hybrids, and applications," *IEEE Access*, vol. 8, pp. 77746–77764, 2020.
[34] I. I. M. Manhrawy, M. Qaraad, and P. El-Kafrawy, "Hybrid feature selection model based on relief-based algorithms and regulizer algorithms for cancer classification," *Concurrency Comput. Pract. Exper.*, vol. 33, no. 17, p. e6200, Sep. 2021.
[35] (2020). *Br35H: Brain Tumor Detection*. [Online]. Available: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection
[36] (2022). *Endometrial—Datasets—PLCO—The Cancer Data Access System*. [Online]. Available: https://cdas.cancer.gov/datasets/plco/17/#: :text=The%20Endometrial%20dataset%20is%20a
[37] I. Danihelka, B. Lakshminarayanan, B. Uria, D. Wierstra, and P. Dayan, "Comparison of maximum likelihood and GAN-based training of real NVPs," 2017, *arXiv:1705.05263*.
[38] N. Chen and C. Li, "Hyperspectral image classification approach based on Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Dec. 2020, pp. 53–63.
[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[40] F. J. Moreno-Barea, J. M. Jerez, and L. Franco, "GAN-based data augmentation for prediction improvement using gene expression data in cancer," in *Proc. 22nd Int. Conf.*, 2022, pp. 28–42.
[41] P. McDaniel, N. Papernot, and Z. B. Celik, "Machine learning in adversarial settings," *IEEE Secur. Privacy*, vol. 14, no. 3, pp. 68–72, May 2016.
[42] H. Fathi, H. AlSalman, A. Gumaei, I. I. M. Manhrawy, A. G. Hussien, and P. El-Kafrawy, "An efficient cancer classification model using microarray and high-dimensional data," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–14, Dec. 2021.
[43] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine learning proceedings*, San Mateo, CA, USA: Morgan Kaufmann, 1994, pp. 121–129.
[44] A. Mobiny, S. Moulik, N. Garg, C. C. Wu, and H. V. Nguyen, "Capsule networks for lung cancer screening," in *Lung Cancer and Imaging*. Bristol, U.K.: IOP Publishing, 2019, pp. 1–2.

U. RAVINDRAN received the B.E. degree in CSE from Anna University, Chennai, India, and the M.Tech. degree in CSE from the Bharath Institute of Higher Education and Research, Chennai. He is currently pursuing the Ph.D. degree with the School of Computer Science and Information Systems, Vellore Institute of Technology, Vellore, India. He has around 12 years of teaching experience. His research interests include machine learning and deep learning.

C. GUNAVATHI received the B.E. degree in CSE from Bharathidasan University, and the M.E. degree in CSE and the Ph.D. degree from Anna University, Chennai, India. She is currently a Professor with the School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. She has around 17 years of teaching experience. Her research interests include data mining, bioinformatics, and soft computing techniques.