STM JOURNALS

International Journal of
Genetic Modifications and Recombinations

Volume 1, Issue 2, 2023
DOI (Journal): 10.37591/IJGMR

https://journals.stmjournals.com/ijgmr/

Research IJGMR

Dimensionality Reduction Techniques and their Applications in Cancer Classification: A Comprehensive Review

Abrar Yaqoob[1,*], Mohd Abas Bhat[2], Zeba Khan[3]

Abstract

Dimensionality reduction techniques have become a vital tool in the investigation of high-dimensional data like gene expression profiles in cancer research. Here is a review, we deliver a comprehensive overview of dimensionality reduction techniques and their applications in cancer classification. Firstly, we introduce the concepts and approaches of dimensionality reduction, and after that, we explore several methods for decreasing dimensionality. These techniques include Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), and t-Distributed Stochastic Neighbour Embedding (t-SNE). We then present a comprehensive review of applications of these techniques in cancer classification, counting lung cancer, colon cancer, breast cancer, and leukemia. Moreover, we discuss the advantages and disadvantages of different dimensionality reduction techniques in cancer classification, as well as their limitations and future directions. Finally, we summarize the most recent stage in the area and make it available for use of some recommendations for future studies. Overall, this review highlights the importance of dimensionality reduction techniques in classification of cancer and provides a valuable resource for researchers working in this field.

Keywords: Feature extraction, cancer classification, dimensionality reduction, feature selection

INTRODUCTION

A massive portion of digital data has really been continuously produced during the last few years in a variety of application areas. Moreover, data are becoming exponentially larger, heterogeneous, complicated, and dimensional. Applications for High Dimensional Data (HDD) have been found in a variety of industries, such as education, biomedicine, the web, health, social media, and business. The enormous capacity of fresh High Dimensional Data is constantly changing and evolving in many formats. The enormous dimensionality of the data might provide several challenges for machine learning models in terms of precise categorization, visualization, and pattern detection. Due to the enormous computational complexity required, learning in contexts with high-dimensional data or a large number of characteristics can become problematic [1]. Dimensionality reduction refers to the process of converting highly dimensional data into a meaningful representation with decreasing dimensionality. Idealized, the reduced representation's dimensionality should coincide with the data's natural dimensionality. The intrinsic dimension of data is the smallest set of parameters required to describe the data's observable properties [2]. Because it addresses the detrimental impacts of dimensionality and other characteristics linked to

*Author for Correspondence
Abrar Yaqoob
E-mail: abraryaqoob77@gmail.com

[1]Research Scholar, Department of Mathematics VIT Bhopal, Sehore, Madhya Pradesh, India
[2]Student, Department of Economics, Kashmir University, Srinagar, India
[3]Research Scholar Department of biotech, vit university Bhopal, Sehore, Madhya Pradesh, India

Received Date: September 13, 2023
Accepted Date: September 29, 2023
Published Date: October 25, 2023

Citation: Abrar Yaqoob, Mohd Abas Bhat, Zeba Khan. Dimensionality Reduction Techniques and their Applications in Cancer Classification: A Comprehensive Review. International Journal of Genetic Modifications and Recombinations. 2023;1(2): 34–45p.

high-dimensional data, the process of dimensionality reduction holds significance across numerous industries. Therefore, the classification, dimensionality reduction facilities, visualization, compression of high-dimensional data, and among other things [3].

Numerous research articles on multivariate statistics for microarray data processing have been published in the recent several years in the area of statistics, bioinformatics, ML, and computational biology. High-dimensional microarray data have been explored in relation to the majority of the standard multivariate statistical problems [4]. Data analysis is mostly required for the following purposes in biomedical applications:

- Gene selection is a feature selection process that identifies genes that are significantly related with a certain class [5].
- Classification: based on gene expression patterns, categorising illnesses or forecasting outcomes, and maybe determining the optimal treatment for a specific genetic profile [6].
- Clustering: process of discovering new biological classifications or improving existing ones [7].

Clustering can be employed to uncover groupings of genes that are expressed similarly in the hope of discovering that they together serve the same function. Another issue of interest is the categorization of microarray data for illness prediction, such as cancer, utilising gene expression levels [8]. Dimension reduction and classifier design are required for the Samples of gene expression data are classified. Thus, dimension reduction is a critical technique for categorization in order to appropriately analyse gene expression patterns [9]. The aim of classifying cancer microarray data is to construct a model that efficiently and accurately differentiates gene expressions within samples, i.e. categorise tissue samples into various tumour classifications. The most well-known algorithms for classification include nearest neighbour classification, Bayesian, Decision tree, Random forest methods, ANN and SVM [10].

In the recent past, several microarray studies have focused on the use of gene expression patterns for cancer diagnosis. In the literature, many selection of gene approaches and classification algorithms are presented that can minimise dimensionality by deleting unnecessary, noisy gene and redundant for accurate cancer classification [11].

Cancer classification refers to the process of identifying the type of cancer a patient has based on the characteristics of the tumor cells. Accurate cancer classification is critical for determining the most effective treatment plan and improving patient outcomes. However, Cancer categorization is a difficult process since there are so many different characteristics, such as gene expression levels, protein expression, and other biological factors. One solution to address these challenges is to use dimensionality reduction algorithms to decrease the quantity of variables used in the classification of cancer. Techniques for dimension reduction can be used to extract the most crucial aspects from high-dimensional data, enabling a more accurate and efficient analysis of the data [12].

This review paper's objective is to present a thorough overview of dimensionality reduction methods and how they are used in the categorization of cancer. We will discuss the challenges of cancer classification and how dimensionality reduction techniques can be used to address these challenges. Additionally, we will review the various dimensionality reduction techniques used in cancer classification, their advantages and limitations, and provide case studies to illustrate their application. Ultimately, this review paper aims to provide insights into the use of dimensionality reduction techniques in cancer classification and provide recommendations for future research.

MICROARRAY GENE EXPRESSION DATA ANALYSIS CHALLENGES

Microarray technique enabled investigators to quantify the appearance of hundreds of genes in a just one experiment. Microarray data is characterized by a restricted sample size and elevated dimensionality. The quantity of variables (genes) vastly out numbers the number of samples n in gene expression microarray data, which is known as the "Dimensionality’s cruse" problem [13]. Figure 1

displays the processing of microarray gene expression data**.** Dimension reduction is critical in DNA microarray analysis to avoid the "Dimensionality's cruse". The scientific community receives a massive quantity of data from microarray tests, but without the right methodology and tools, important information and knowledge might remain concealed. Analytical and statistical difficulties result from the enormous volume of raw gene expression data. The nature of the microarray data presents a difficulty to statisticians [14]. The entire number of potential gene combinations will have a significant impact on the most effective statistical model. Therefore, data mining and statistical analysis will play a significant role in how the microarray technology affects biology. Because microarray data has a high amount of complexity but a limited number of patterns, conventional statistical approaches provide incorrect results. As a result, technologies capable of processing and exploring massive data volumes are required. For the analysis of massive data sets, the disciplines of data mining and machine learning provide a wide range of methods and resources. To build correlations within the obtained array data and permit meaningful classification, a sophisticated data mining and analytical solution is required [15].

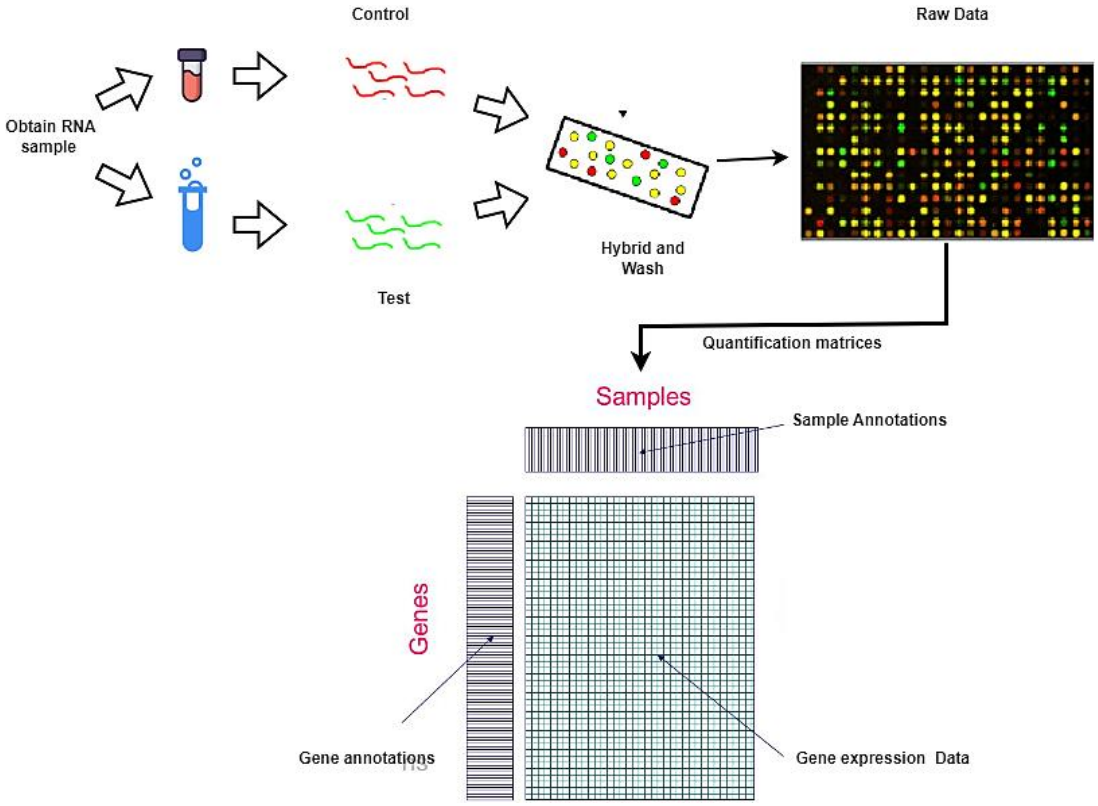

Figure 1. Formation of microarray gene expression data.

Exploration of information using ML is the process of employing various automatic or semi-automatic approaches to extract meaningful, non-trivial and possibly useful knowledge (trends, patterns, correlations, anomalies, rules and dependencies) from a huge quantity of data. Data mining is arguably finest understood as a process that encompasses a larger variety of integrated procedures and tools than standard statistical approaches, includes databases, machine learning, knowledge-based approaches, network technologies, modelling, algorithms, and uncertainty handling [16].

A DNA microarray's gene expression data, which depict the molecular state of a cell, has considerable potential as a diagnostic tool. Regression, clustering, association, Discriminant analysis, and deviation detection are examples of common microarray data mining analyses. Different research has used a variety of machine learning methods to analyse microarray gene expression data, some examples are k-Nearest Neighbors, Artificial Neural Networks, Naive Bayes, Genetic Algorithms, Bayesian Networks,

Decision Trees, Rough Sets, Emerging Patterns, and Self-Organizing Maps are examples of techniques. [17]. Given the enormous number of genes involved in categorization, the training data sets that are currently available often have a rather limited sample size. Theoretically, larger genes should have greater discriminatory power, but in practise, learning is greatly slowed down by huge genes. As well as impairing model simplicity and causing the classifier to over fit the training data. It is possible to successfully extract the genes that directly affect categorization using dimension reduction. In this article, we emphasise dimension reduction and picking out of potentially significant genes for the molecular categorization of malignancy using common machine learning approaches [18].

DIMENSIONALITY REDUCTION

The fundamental challenge with most machine learning algorithm for classifying microarray data is having to train with a high quantity of genes. In order to build an accurate characterisation of the classification problem, a learning algorithm is often given a large number of candidate features (genes). Since 10 years ago. The count of variables or features utilized in machine learning and pattern recognition applications has escalated from tens to hundreds. To address the issue of removing unnecessary and duplicate characteristics that are a burden for various demanding jobs, a variety of machine learning algorithms have been created [19].

Dimensionality reduction is the technique of minimising information loss while diminish the quantity of features or variables within a dataset. Given that working with high-dimensional datasets may be challenging and they may experience the "curse of dimensionality," it is a vital approach in data pre-processing and machine learning [20]. Figure 2 gives an Overview of Dimensionality Reduction Approaches.

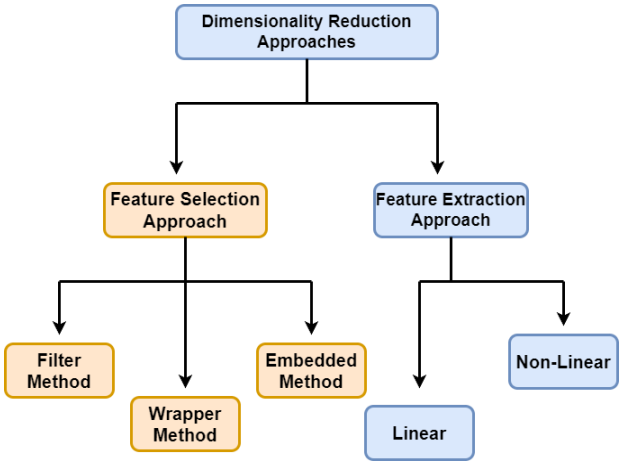

Figure 2. Overview of Dimensionality Reduction Approaches.

Dimensionality reduction approaches are classified into two types:

Feature Selection

The procedure of reducing the dimensionality of a dataset by finding a collection of qualities that accurately describe the data is known as feature selection. This process involves relevant feature and selection important while removing redundant and inappropriate one. The purpose is to create the least possible subset of features but still accurately represents the entire input data. There are many benefits to feature selection, including reduced data size, decreased storage requirements, improved prediction accuracy, avoidance of overfitting, and reduced execution and training times [21]. There are two stages to the algorithmic step of feature selection: Subset Generation and Subset Evaluation. Subset Generation involves generating subsets from the input dataset, while Subset Evaluation is used to determine if the generated subset is optimal. See "Figure 3" for an illustration of the overall feature selection process [22, 23].

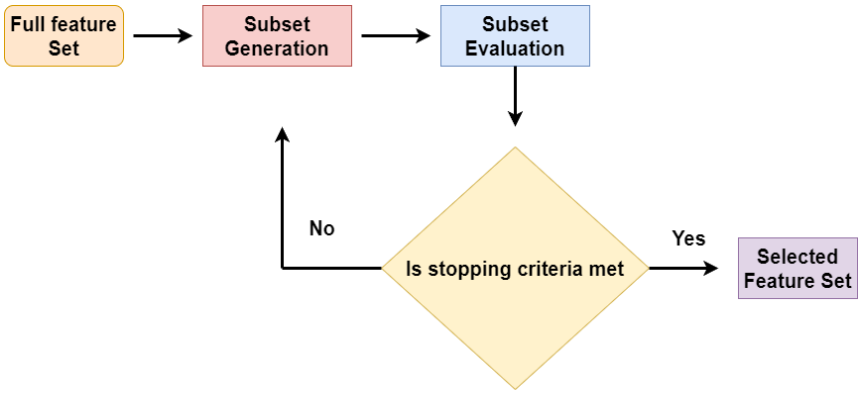

Figure 3. Feature selection process.

Feature Selection Methods

The objective of feature selection involves opting for a subset of features from a larger collection of options, based on their relevance and redundancy. There are several methods for evaluating the usefulness of features, wrapper, embedded, including filter, and hybrid techniques. More recently, a new method called ensemble feature selection has emerged [24].

Filter Method

The filter method is a feature selection algorithm that evaluates the inherent properties of features prior to the learning tasks (Figure 4). It uses four measurement criteria, including information, dependency, consistency, and distance to measure the feature characteristics. This method is independent of the data mining algorithm and uses statistical standards to evaluate the ranking of the subset. It is known for its good performance and high-efficiency computing, scalability in high-dimensional datasets, and outperforming the wrapper technique. However, a downside of this method is that it does not take into account the integration of the selected subset and the performance of the induction process. [25]. Table 1 gives Pros and Cons of Filter Method.

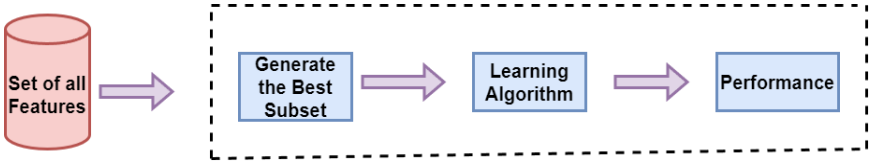

Figure 4. Workflow of filter method.

Table 1. Pros and Cons of Filter Method

Model search	Advantages	Disadvantages	Examples
	Fast	Ignores feature dependencies	$\chi 2$
Filter Method	Scalable Independent of the classifier.	Some features which as a group have strong discriminatory power but are weak as individual features will be ignored	Euclidean distance
	The models feature dependencies.	Slower than univariate techniques	t-test
	Independent of the classifier.	Features are considered independently.	Information gain
	Better computational complexity than wrapper methods	Less scalable than univariate techniques.	Gain ratio

The relevance of a feature in relation to the data or output is an important consideration, and various definitions and measurements have been proposed in the literature. A feature can be considered unimportant if its relationship with the class labels is conditionally independent, according to one usable definition. This means that a feature may not have any impact on the input data, but it must have an impact on the class labels to be considered relevant. Inter-feature correlation is also important for

determining unique features. However, the underlying distribution in practical applications is often unknown and is measured by classifier accuracy. As a result, there may be different sets of features that achieve the same classifier accuracy, leading to non-uniqueness in optimal feature subsets [19].

Wrapper Method

Wrapper approaches, on the other hand, employ an inductive ML algorithm to estimate the worth of a given subset or collection of qualities (Figure 5). This approach is often considered a more favourable option in scenarios involving supervised learning due to its advantages by interacting with the inductive algorithm to evaluate possibilities, they accommodate for the algorithm's specific biases. If a trait is conditionally independent of the class labels, it is considered irrelevant, according to one usable definition. However, even for somewhat sophisticated algorithms, the amount of executions necessary during feature search may result in a substantial computational cost, particularly as we move toward more exhaustive search techniques [26]. Table 2 shows Pros and Cons of Wrapper method.

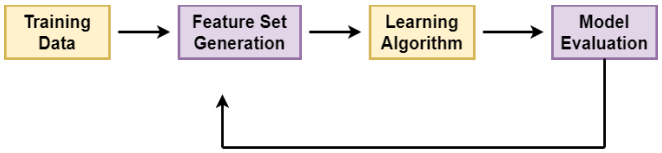

Figure 5. Working procedure of wrapper method.

Table 2. Pros and cons of wrapper method.

Model Search	Advantages	Disadvantages	Examples
	Simple Interacts with the classifier	Risk of over fitting	Sequential forward selection (SFS)
Wrapper Method	Small overfitting risk	More prone than randomized	Sequential backward elimination (SBE)
	Less computationally	Computationally intensive	
	Prone to local optima	Discriminative power	Beam search
	Consider the dependence among features	Lower shorter training times	Simulated annealing
	Less prone to local optima	Classifier dependent selection	Randomized hill climbing
	Interacts with the classifier	Higher risk of over-fitting than deterministic algorithms	Genetic algorithms
	Models feature dependencies		Ant Colony Optimization
	Higher performance accuracy than filter		Rough set methods

Embedded Method

Embedded feature selection approaches combine the feature selection and model training processes, allowing for simultaneous optimisation of both (Figure 6). These strategies try to locate the most important characteristics for the model during the training phase, minimising data dimensionality and boosting model performance [27]. Embedded approaches are excellent in identifying the most important model characteristics and improving model performance. However, keep in mind that these strategies may not be appropriate for all datasets and model types. Table 3 shows Pros and Cons of Embedded Method.

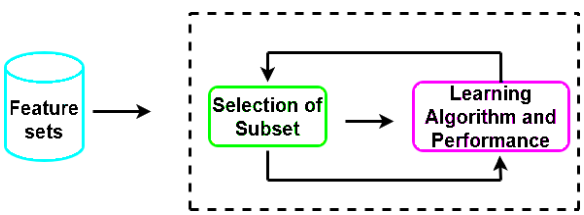

Figure 6. Overview of embedded method.

Table 3. Pros and cons of embedded method.

Model Search	Advantages	Disadvantages	Examples
	Interacts with the classifier	Classifier dependent selection	Decision trees
Embedded Method	The models feature dependencies better	Consider the dependence among features	Weighted naive Bayes
	computational complexity than the wrapper		Feature selection using the weight vector of SVM
	Higher performance, accuracy than filter		Random forests
	Less prone to over-fitting than wrapper		Least absolute shrinkage and selection operator (LASSO)
	Preserving data characteristics for interpretability		

Hybrid Technique Method

A hybrid technique selection of feature, combines multiple selection of feature methods to select a subset of the most applicable feature for a given task or problem. This approach seeks to address the limitations and biases of individual feature selection methods by leveraging their strengths and complementing their weaknesses [28].

Feature Extraction

The technique of collecting discriminating data from a set of samples is called feature extraction. For the extraction of medically useful information from the textures, features must be computed. The traits, which may not be visually visible but are relevant to the diagnostic issue, can be thought of as supplements to the researchers' visual abilities. Effective and distinctive features are extracted using a variety of feature extraction techniques. Below is an explanation of a few features extraction techniques [29].

In order to extract valuable information from pictures, many feature extraction approaches are utilised during the processing of medical images.

Gray-Level Co-Occurrence Matrix (GLCM)

The GLCM texture analysis approach is commonly utilised in medical image processing. It involves computing the probability of co-occurrence of pixel values at specific pixel distances and directions in an image. The co-occurrence matrix is then used to compute various statistical measures such as contrast, correlation, energy, and homogeneity. These metrics can be utilised for classification or segmentation tasks as well as features to characterise the texture of a picture [30].

Local Binary Patterns (LBP)

LBP is a straightforward yet powerful tool for analysing texture. Each pixel in an image is compared to its neighbouring pixels, and a binary value is allocated on the basis of whether the neighbouring pixels have greater or lower values than the centre pixel. This method is repeated for each pixel in the image to form a binary pattern. These patterns can then be used as features to describe the texture of an image [31].

Gabor Wavelets

Gabor wavelets are a type of filter that is used to analyse the frequency and orientation content of an image. They are particularly useful for analysing texture because they can capture both the fine and coarse details of an image. Gabor wavelets can be used to extract features such as mean amplitude, mean frequency, and classification or segmentation task-specific orientation [32].

Histogram of Oriented Gradients (HOG)

HOG is a feature extraction techniques that involves computing the gradient magnitude and direction of an image and then grouping these gradients into histograms based on their orientation. These

histograms can then be used as features to describe the texture of an image. HOG has been shown to be particularly effective for object detection and recognition tasks [33].

Convolutional Neural Networks (CNN)

CNNs are a type of DL (Deep Learning) method that can be used for feature extraction in medical image processing. Their purpose is to acquire and derive features from pictures automatically by employing convolutional layers, which apply filters to an image to extract certain information. The output from the convolutional layers can then be used as features for classification or segmentation tasks. CNNs have been demonstrated to be extremely successful for several medical image processing applications, as well as tumour identification and segmentation [34].

Linear and Non-Linear Approaches of Feature Extraction

Dimensionality reduction is an approach used to decrease the variables within a dataset, all while retaining crucial connections and informational elements. There are two basic categories for this reduction process: linear techniques and non-linear approaches [35].

The linear dimensionality reduction approach involves generating a new set of variables by combining the old variables in linear fashion. The purpose is to encompass the primary information from the original data by pinpointing the dimensions that exhibit the highest variability. Principal component analysis (PCA) is a popular linear approach. Conversely, non-linear dimensionality reduction methods use non-linear transformations of the original variables to generate a new set of variables that capture the essential information in the data [36].

Both linear and non-linear dimensionality reduction methods have their strengths and weaknesses. The selection of a method depends on the specific circumstances and the nature of the data being analysed. Since it aids in reducing noise, improving visualization, and boosting the performance of other ML algorithms, Dimensionality reduction is an vital element in many data and ML analysis applications [37].

Applications of Dimensionality Reduction in cancer classification

The principal component analysis (PCA) dimension reduction technique presented by Adiwijaya et al. includes calculating the variance proportion for eigenvector selection Levenberg-Marquardt Backpropagation (LMBP) and (SVM) Support Vector Machine were chosen as the classification methods. According to the testing, the LMBP classification approach was more reliable than SVM. The average accuracy of the LMBP technique was 96.07%, while the SVM's accuracy was 94.98% [4].

In order to minimise feature dimensions, Kai-Lin Tang et al. suggested an approach grounded on statistical moments. SELDI-TOF data were separated into multiple intervals after revision and t-testing. For each interval, the average, variability, asymmetry, and peakedness of four statistical moments were determined and utilised as representative variables. Thus, the data's large dimensionality may be quickly decreased. The data were also utilised in kernel PLS models to enhance effectiveness and classification performance. The approach has a mean sensitivity of 0.9950, a mean specificity of 0.9916, an average accuracy of 0.9935, and a correlation value of 0.9869 after 100 rounds of five-fold cross-validations [38].

According to a technique put out by Md. Faisal Kabir et al., Both kernel-based PCA and auto-encoder approaches can be employed to decrease the dimensionality of RNA sequencing data. Following that, the original dataset, its dimensionally reduced counterpart, and the relevant cancer-related data were used to train and evaluate two machine learning classifiers: a neural network and a support vector machine. The performance of the classifiers was assessed using a range of metrics, encompassing accuracy, precision, recall, F-Measure, receiver operating characteristic curve, and area under the curve. The outcomes demonstrated that dimensionality reduction has a favourable impact on the cross-validation performance of classifiers [39].

Using knowledge-based categorization, Mehrbakhsh Nilashi et al. created a method for identifying breast cancer. The system makes use of algorithms for clustering, removing noise, and classifying data. To group comparable data, Expectation Maximisation (EM) clustering is used. In the knowledge-based approach, fuzzy rules for breast cancer categorization are produced using categorization and Regression Trees (CART). Multi-collinearity is addressed by incorporating Principal Component Analysis (PCA). The experimental outcomes indicate a substaintial enhancement in the accuracy of breast cancer prediction by incorporating the Wisconsin Diagnostic Breast Cancer and Mammographic Mass databases [40].

Sarah M. Ayyad et al. propose a new gene expression data classification technique known as Modified k-nearest neighbor (MKNN). MKNN consists of two implementations: The two types of modified KNNs are smallest and biggest. Both implementations aim to improve the performance of KNN by expending a novel weighting plan to select robust neighbors from the training data. The researchers conducted experiments on six gene expression datasets and found that MKNN outperforms traditional and recent techniques in terms of classification accuracy, precision, and recall. Additionally, Comparing testing times for KNN and weighted KNN, MKNN is faster. [41].

Hanaa Salem and colleagues developed a unique strategy to categorise human cancer illnesses based on gene expression patterns. Their method blends Information Gain (IG) with the Standard Genetic Algorithm (SGA) to efficiently select and minimise features. The evaluation of the system encompassed seven cancer datasets and involved a comparison with alternative methodologies, showing that Genetic Algorithm generally improves classification performance [42].

M. Dashtban et al. suggested a new strategy for finding predictive genes for cancer categorization that combines genetic algorithms and artificial intelligence. They reduced the dimensionality of the features using a filter approach, and they then used a genetic algorithm with integer coding, a genotype with variable length, adaptive parameter tuning, and improved operators. The algorithm's behaviour, encompassing trends in convergence, variations in mutation and crossover rates, and the duration of execution, was investigated and found to be consistent with earlier studies. They examined two filter approaches, Fisher score and, taking into account similarities, gene quality, and the influence on the evolutionary strategy. Statistical tests were performed to examine the choice of classifier, dataset, and filter technique, and significant variations in performance were discovered. The top genes discovered using the recommended method were published after it was evaluated on five high-dimensional cancer datasets. In contrast to cutting-edge techniques, the suggested strategy outperformed earlier methods in the DLBCL dataset. [43].

Sherif et al. developed a classification model aimed at identifying cervical cancer utilizing a set of risk factors. For the purpose of boosting the model’s performance, they combined two techniques for feature reduction: principal analysis (PCA) and recursive feature elimention. They also incorporated the synthetic minority oversampling technique (SMOTE) to address the prevalent issue of class imbalance often encountered in medical datasets. The classification approach they employed was Random Forest (RF). The dataset for their research encompassed 32 risk factors and four target variables: Hinselmann, Schiller, Cytology, and Biopsy. The outcomes of their comparative analysis revealed that the most effective enhancement was achieved by combining the random forest classification strategy with SMOTE, leading to improved classification performance [44].

Teresa Araujo et al. suggested a technique for classifying breast biopsy pictures stained with hematoxylin and eosin using Convolutional Neural Networks (CNNs). Normal tissue, benign lesion, in situ cancer, and aggressive carcinoma are the four categories of pictures. Furthermore, they are classified into two types: carcinoma and non-carcinoma. The network architecture collects data at several scales, taking into account both nuclei and general tissue organisation. Because of its architecture, the technique may be used to full histology pictures. The retrieved features from the CNN

are also used to train a Support Vector Machine classifier. The approach obtains 77.8% accuracy for four-class classification and 83.3% accuracy for carcinoma/non-carcinoma classification. Notably, the method's sensitivity for cancer cases is 95.6%. [45].

Rui Yan et al. anticipated a unique hybrid deep neural network architecture for the categorization of histological pictures of breast cancer. Their solution combines the strengths of convolutional and recurrent neural networks to exploit the picture patches' rich multilayer characteristics. The model effectively retains both spatial correlations, both immediate and extended, among patches. by integrating these networks. The results of thorough experimentation show that their suggested method outperforms the present state-of-the-art methodology, with an amazing average accuracy of 91.3% in the 4-class classification challenge [46].

CONCLUSION

The review paper provides an in-depth analysis of dimensionality reduction techniques and their importance in cancer classification. It covers various aspects, including the concepts and approaches of dimensionality reduction, different techniques used in cancer classification, and their applications in various cancer types. Additionally, the report examines the benefits and drawbacks of each approach, as well as their limits in cancer categorization The review concludes by summarizing the current cutting-edge technology in the field and providing recommendations for future research.

Overall, the review paper emphasizes the significance of dimensionality reduction techniques in cancer research and their potential applications in cancer classification. By providing a comprehensive overview of different techniques, the review serves as an essential resource for researchers working in this field. It highlights the need for further research in developing more efficient and accurate dimensionality reduction techniques and their integration with other machine learning approaches in cancer classification. The review contributes significantly to the advancement of cancer research and provides valuable insights for future directions in this field.

REFERENCES

1. R. Aziz, C. K. Verma, and N. Srivastava, “Artificial Neural Network Classification of High Dimensional Data with Novel Optimization Approach of Dimension Reduction,” *Ann. Data Sci.*, vol. 5, no. 4, pp. 615–635, Dec. 2018, doi: 10.1007/s40745-018-0155-2.
2. S. Ayesha, M. K. Hanif, and R. Talib, “Overview and comparative study of dimensionality reduction techniques for high dimensional data,” *Inf. Fusion*, vol. 59, no. January, pp. 44–58, 2020, doi: 10.1016/j.inffus.2020.01.005.
3. L. J. P. Van Der Maaten, E. O. Postma, and H. J. Van Den Herik, “Dimensionality Reduction: A Comparative Review,” *J. Mach. Learn. Res.*, vol. 10, pp. 1–41, 2009, doi: 10.1080/13506280444000102.
4. U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo, “Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification,” no. November, 2018, doi: 10.3844/jcssp.2018.1521.1530.
5. I. Guyon, “Gene Selection for Cancer Classification,” pp. 389–422, 2002.
6. H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, “A hybrid feature selection algorithm for gene expression data classification,” *Neurocomputing*, vol. 256, pp. 56–62, Sep. 2017, doi: 10.1016/j.neucom.2016.07.080.
7. F. Murtagh, “A survey of recent advances in hierarchical clustering algorithms,” *Comput. J.*, vol. 26, no. 4, pp. 354–359, 1983, doi: 10.1093/comjnl/26.4.354.
8. S. Shukla and S. Naganna, “A Review ON K-means DATA Clustering APPROACH,” vol. 4, no. 17, pp. 1847–1860, 2014.
9. A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, “Benchmark for filter methods for feature selection in high-dimensional classification data,” *Comput. Stat. Data Anal.*, vol. 143, p. 106839, 2020, doi: 10.1016/j.csda.2019.106839.

10. M. Jansi Rani and D. Devaraj, "Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification," *J. Med. Syst.*, vol. 43, no. 8, Aug. 2019, doi: 10.1007/s10916-019-1372-8.
11. O. Ahmad Alomari, A. Tajudin Khader, M. Azmi Al-Betar, and L. Mohammad Abualigah, "Gene selection for cancer classification by combining minimum redundancy maximum relevancy and bat-inspired algorithm," *Int. J. Data Min. Bioinform.*, vol. 19, no. 1, pp. 32–51, 2017, doi: 10.1504/IJDMB.2017.088538.
12. J. Galon *et al.*, "Cancer classification using the Immunoscore: A worldwide task force," *J. Transl. Med.*, vol. 10, no. 1, 2012, doi: 10.1186/1479-5876-10-205.
13. N. Almugren and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE Access*, vol. 7, pp. 78533–78548, 2019, doi: 10.1109/ACCESS.2019.2922987.
14. V. Elyasigomari, D. A. Lee, H. R. C. Screen, and M. H. Shaheed, "Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification," *J. Biomed. Inform.*, vol. 67, pp. 11–20, 2017, doi: 10.1016/j.jbi.2017.01.016.
15. I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification," *Appl. Soft Comput.*, vol. 62, pp. 203–215, 2018, doi: 10.1016/j.asoc.2017.09.038.
16. G. Nguyen *et al.*, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77–124, 2019, doi: 10.1007/s10462-018-09679-z.
17. S. Cho and H. Won, "Machine Learning in DNA Microarray Analysis for Cancer Classification," no. May 2014, 2018.
18. N. Almugren and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE Access*, vol. 7. Institute of Electrical and Electronics Engineers Inc., pp. 78533–78548, 2019. doi: 10.1109/ACCESS.2019.2922987.
19. A. Yaqoob, R. M. Aziz, N. K. Verma, P. Lalwani, and A. Makrariya, "A Review on Nature-Inspired Algorithms for Cancer Disease Prediction and Classification," 2023.
20. A. Ghodsi, "Dimensionality Reduction A Short Tutorial."
21. A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, no. May 2015, pp. 1200–1205, 2015, doi: 10.1109/MIPRO.2015.7160458.
22. D. A. A. Gnana, "Literature Review on Feature Selection Methods for High-Dimensional Data Literature Review on Feature Selection Methods for High-Dimensional Data," no. August, 2016, doi: 10.5120/ijca2016908317.
23. C. C. Aggarwal, "Educational and software resources for data classification," *Data Classif. Algorithms Appl.*, pp. 657–665, 2014, doi: 10.1201/b17320.
24. V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 483–519, 2013, doi: 10.1007/s10115-012-0487-8.
25. G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.
26. B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Comput. Biol. Med.*, vol. 112, no. July, p. 103375, 2019, doi: 10.1016/j.compbiomed.2019.103375.
27. H. J. Ferreau *et al.*, "Embedded Optimization Methods for Industrial Automatic Control," vol. 50, no. 1, pp. 13194–13209, 2017, doi: 10.1016/j.ifacol.2017.08.1946.
28. I. Gitchat, "特征选择（Feature Selection）特征选择 Feature Selection," *Comput. Vis.*, vol. 392, no. March, pp. 1–10, 2018, [Online]. Available: http://link.springer.com/10.1007/978-3-030-03243-2_299-1
29. P. Lamba and K. Rawal, "A Survey of Algorithms for Feature Extraction and Feature Classification Methods."

30. F. Roberti de Siqueira, W. Robson Schwartz, and H. Pedrini, "Multi-scale gray level co-occurrence matrices for texture description," *Neurocomputing*, vol. 120, pp. 336–345, 2013, doi: 10.1016/j.neucom.2012.09.042.
31. A. Hadid, J. Ylioinas, M. Bengherabi, M. Ghahramani, and A. Taleb-Ahmed, "Gender and texture classification: A comparative analysis using 13 variants of local binary patterns," *Pattern Recognit. Lett.*, vol. 68, pp. 231–238, 2015, doi: 10.1016/j.patrec.2015.04.017.
32. Á. Serrano, I. M. de Diego, C. Conde, and E. Cabello, "Recent advances in face biometrics with Gabor wavelets: A review," *Pattern Recognit. Lett.*, vol. 31, no. 5, pp. 372–381, 2010, doi: 10.1016/j.patrec.2009.11.002.
33. S. E. Lee, K. Min, and T. Suh, "Accelerating Histograms of Oriented Gradients descriptor extraction for pedestrian recognition," *Comput. Electr. Eng.*, vol. 39, no. 4, pp. 1043–1048, 2013, doi: 10.1016/j.compeleceng.2013.04.001.
34. N. Aloysius, A. V. Vidyapeetham, G. Madathilkulangara, and A. V. Vidyapeetham, "A Review on Deep Convolutional Neural Networks," no. April, 2017, doi: 10.1109/ICCSP.2017.8286426.
35. I. Guyon, S. Gunn, and M. Nikravesh, "Feature Extraction," 2006.
36. J. Behmann, A. Mahlein, T. Rumpf, C. Ro, and L. Plu, "A review of advanced machine learning methods for the detection of biotic stress in precision crop protection," pp. 239–260, 2015, doi: 10.1007/s11119-014-9372-7.
37. K. K. Kumar, K. Chaduvula, and B. R. Markapudi, "A Detailed Survey On Feature Extraction Techniques In Image Processing For Medical Image Analysis," vol. 07, no. 10, pp. 2275–2284, 2020.
38. K. L. Tang, T. H. Li, W. W. Xiong, and K. Chen, "Ovarian cancer classification based on dimensionality reduction for SELDI-TOF data," *BMC Bioinformatics*, vol. 11, 2010, doi: 10.1186/1471-2105-11-109.
39. M. F. Kabir, T. Chen, and S. A. Ludwig, "A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction," *Healthc. Anal.*, vol. 3, no. November 2022, p. 100125, 2023, doi: 10.1016/j.health.2022.100125.
40. M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telemat. Informatics*, vol. 34, no. 4, pp. 133–144, 2017, doi: 10.1016/j.tele.2017.01.007.
41. S. M. Ayyad, A. I. Saleh, and L. M. Labib, "Gene expression cancer classification using modified K-Nearest Neighbors technique," *BioSystems*, vol. 176, no. January, pp. 41–51, 2019, doi: 10.1016/j.biosystems.2018.12.009.
42. H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Appl. Soft Comput. J.*, vol. 50, pp. 124–134, 2017, doi: 10.1016/j.asoc.2016.11.026.
43. M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, 2017, doi: 10.1016/j.ygeno.2017.01.004.
44. S. F. Abdoh, M. Abo Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," *IEEE Access*, vol. 6, pp. 59475–59485, 2018, doi: 10.1109/ACCESS.2018.2874063.
45. A. Madduri, S. S. Adusumalli, H. S. Katragadda, M. K. R. Dontireddy, and P. S. Suhasini, "Classification of Breast Cancer Histopathological Images using Convolutional Neural Networks," *Proc. 8th Int. Conf. Signal Process. Integr. Networks, SPIN 2021*, pp. 755–759, 2021, doi: 10.1109/SPIN52536.2021.9566015.
46. R. Yan *et al.*, "Breast cancer histopathological image classification using a hybrid deep neural network," *Methods*, vol. 173, no. June 2019, pp. 52–60, 2020, doi: 10.1016/j.ymeth.2019.06.014.

