processes

MDPI

*Review*

# A Review on Recent Progress in Machine Learning and Deep Learning Methods for Cancer Classification on Gene Expression Data

**Aina Umairah Mazlan [1], Noor Azida Sahabudin [1,*], Muhammad Akmal Remli [2,3,*], Nor Syahidatul Nadiah Ismail [1], Mohd Saberi Mohamad [4], Hui Wen Nies [5] and Nor Bakiah Abd Warif [6]**

1 Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Pekan 26600, Pahang, Malaysia; ainaumairahm@gmail.com (A.U.M.); nadiahismail@ump.edu.my (N.S.N.I.)
2 Institute for Artificial Intelligence and Big Data, City Campus, Pengkalan Chepa, Universiti Malaysia Kelantan, Kota Bharu 16100, Kelantan, Malaysia
3 Department of Data Science, City Campus, Universiti Malaysia Kelantan, Pengkalan Chepa, Kota Bharu 16100, Kelantan, Malaysia
4 Health Data Science Lab, Department of Genetics and Genomics, College of Medical and Health Sciences, United Arab Emirates University, AI Ain P.O. Box 17666, United Arab Emirates; saberi@uaeu.ac.ae
5 Artificial Intelligence and Bioinformatics Research Group, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia; huiwennies@utm.my
6 Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400, Johor, Malaysia; norbakiah@uthm.edu.my
* Correspondence: azida@ump.edu.my (N.A.S.); akmal@umk.edu.my (M.A.R.)



**Citation:** Mazlan, A.U.; Sahabudin, N.A.; Remli, M.A.; Ismail, N.S.N.; Mohamad, M.S.; Nies, H.W.; Abd Warif, N.B. A Review on Recent Progress in Machine Learning and Deep Learning Methods for Cancer Classification on Gene Expression Data. *Processes* **2021**, *9*, 1466. https://doi.org/10.3390/pr9081466

Academic Editor: Frederic Cadet

Received: 6 July 2021
Accepted: 18 August 2021
Published: 22 August 2021



**Abstract:** Data-driven model with predictive ability are important to be used in medical and healthcare. However, the most challenging task in predictive modeling is to construct a prediction model, which can be addressed using machine learning (ML) methods. The methods are used to learn and trained the model using a gene expression dataset without being programmed explicitly. Due to the vast amount of gene expression data, this task becomes complex and time consuming. This paper provides a recent review on recent progress in ML and deep learning (DL) for cancer classification, which has received increasing attention in bioinformatics and computational biology. The development of cancer classification methods based on ML and DL is mostly focused on this review. Although many methods have been applied to the cancer classification problem, recent progress shows that most of the successful techniques are those based on supervised and DL methods. In addition, the sources of the healthcare dataset are also described. The development of many machine learning methods for insight analysis in cancer classification has brought a lot of improvement in healthcare. Currently, it seems that there is highly demanded further development of efficient classification methods to address the expansion of healthcare applications.

**Keywords:** machine learning; deep learning; cancer classification; biomarker; gene expression

## 1. Introduction

In the last decades, the production of huge amounts of data is rapidly growing. Machine and computers have become an important aspect of technology in manipulating and extracting meaningful insight into the data. In the medical and healthcare field, a huge number of data has been created using various methods. The data are used in advancing medical operation and breakthrough research. To extract huge amounts of heterogeneous data, research in data mining is demanded. Data mining is the process of discovering a pattern in a dataset [1].

In general, there are three methods of data mining, which are supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the labeled training dataset is used to predict or map input data to the desired output [2]. Under

unsupervised learning, by contrast, no labeled data are given, and the learning algorithm is used to find the meaningful pattern and data distribution, such as clustering. Therefore, the learning model is responsible for identifying patterns or discovering the classes of the data input. In supervised learning, this procedure can be considered a classification problem [3]. The classification task involves a learning process in which the data are categorized into a variety of classes. In unsupervised learning, clustering is a common task where the categories or clusters are searched to describe the data distributions. This approach can be used as a preprocessing task for feature selection.

This paper mainly focuses on discovering recent developments of machine learning (ML) and deep learning (DL) methods for cancer classification. The growth of healthcare data availability and advancement of data analytic tools have led to the enhancement of ML and DL applications in healthcare [4]. ML and DL have shown significant breakthroughs in solving a wide range of scientific problems [5]. In healthcare, AI offers a wide range of applications, including data management, drug research, disease prediction, and design of treatment [6].

## 2. Recent Reviews of Artificial Intelligence Application in Healthcare

Several review papers have been conducted to show the potentials, trends, and future direction of ML and DL applications in genetics, genomics, bioinformatics, and multi-omics studies [7–16]. The significant outcomes of these work mostly contributed to the cancer research field. One review has been conducted to analyze the machine learning applications applied in the genome sequencing data. This caters to the annotation of sequence elements and epigenetic as well as other omics data. The future challenge of machine learning techniques based on supervised, semi-supervised, and unsupervised are also discussed. The authors provided recommendations and guidelines in assisting the best machine learning methods to be used in the analysis of genetic and genomic data [7].

Other reviews and research focused on specific types of cancer, such as prostate cancer [9,17] and epigenetics [12]. The authors reviewed the potential application and machine learning algorithms to be used in prostate cancer and analyzed epigenetics data, respectively. Single-cell RNA sequencing (scRNA-seq) is one of the recent breakthroughs of the large-scale transcriptome profiling of individual single cells in a cell population. The core analysis of the scRNA-seq is to cluster single cells to detect cell subtypes and draw the networks based on the relationships among cells. Unsupervised learning based on clustering such as k-means is reviewed to cluster the scRNA-seq [10]. Work in [11] reviews both clustering and classification methods to measure similarity between scRNA-seq. The authors discuss machine learning and integrated methods and a description of scRNA-seq data.

The combination and fusions of different multi-omics workflow on a single cell level are presented in [13]. The authors discussed how the use of multi-omics workflows could be used to examine cell phenotype and dynamic changes in a metabolomic state. This can accelerate biomarker discovery based on a machine learning approach. A recent study has summarized the important machine learning application and tools in different types of medicine and healthcare. The work also addressed the future directions and challenges in applying ML and automated tools [14]. Another advancement in this field is medical physics. The work in [15] demonstrated the initiative to accelerate the research of AI application of physics applied to healthcare, which is also referred to as medical physics.

## 3. Application of Artificial Intelligence in Modern Healthcare

AI is a branch of computer science discipline that is making significant progress toward applications in various sectors. AI also refers to the development of intelligent computers that are trained to work and act in the same way as people do. AI was used for the growth and enhancement of a wide range of areas and sectors, particularly in healthcare, which allows the machine to learn the data and then make a prediction [18]. Machine learning techniques have been broadly utilized in many applications from assisting healthcare

practitioner tasks including MRI image recognition, genome data analysis, to scientific findings such as classification and prediction. Recent successful applications of AI in healthcare have been made possible by the increased availability of healthcare data and the fast development of big data analysis methods [4].

The advancement of computational power has dramatically changed the landscape of cancer research. The early work back in 2000 has demonstrated the applicability and practicability of the artificial intelligence techniques in healthcare datasets [19]. The use of DNA microarray experiments has generated a large amount of data of gene expression measurements. The important task in gene expression data is to classify the samples into known categories. The data are valuable to be analyzed, finding the hidden pattern, and selecting informative features before building the machine learning model to determine cancer or normal tissue.

There are two types of feature selections in DNA microarray and gene expression data; filter and wrapper gene selection. The filter method is commonly used in the preprocessing step of the data. The step is independent of any AI algorithms. The informative features (we also called 'informative genes') are selected based on statistical approaches. The score from the statistical test such as Pearson's correlation, *t*-test, and ANOVA is used to filter the gene expression data. This can improve the accuracy of cancer classification. The wrapper gene selection, on the other hand, is the approach that uses a subset of features and trains the algorithm using that subset. It is based on the practitioner's inference to add or remove the feature from the subset. However, this method is computationally expensive and also reduced a search problem [20]. More state-of-arts reviews of feature selection techniques can be found in [21].

Machine learning (ML) algorithms, such as the support vector machine (SVM), neural network (NN), and deep learning (DL), are the common popular algorithms used in solving ML problems [4]. AI may use advanced algorithms to learn features from a huge volume of healthcare data and then use insights obtained to help in clinical practice. The ability to learn and be self-correcting in machine learning can be used through boosting algorithms such as XGBoost or AdaBoost to enhance accuracy based on feedback. The boosting algorithm improves the accuracy using an iterative process until the strongest rule is fitted for test observation. Furthermore, an AI may help to reduce diagnostic and therapeutic mistakes in human clinical practice, which are unavoidable. There are two main types of ML techniques: unsupervised learning and supervised learning [22]. AI applications in healthcare commonly use supervised learning methods. In contrast, unsupervised learning is commonly applied for feature reduction or extraction, while supervised learning can be used for predictive modeling. Based on a previous study in [4], the support vector machine (SVM) and neural networks (NN) were the most popular techniques in medical application.

## 4. Cancer Classification with Machine Learning Method

In cancer diagnosis, the classification of a cancerous or non-cancerous gene in different types of cancer plays a vital role in drug discovery [23]. It is important to accurately predict different types of cancer and gene associated with cancer in order to provide better treatment for patients. Classification tasks using ML classifiers allow the machine to distinguish into multiple classes of the entire dataset based on the features correlated with them [3]. The input dataset that is fed into the ML model is normally in the form of numeric data (e.g., gene expression value) or in the form of images (e.g., MRI images) [24].

In the study of cancer classification problems, there are many machine learning methods that have been widely applied. Supervised and unsupervised are now becoming the two often used methods in cancer classification [25]. Unsupervised methods discover the structure of the features of the sample, and one of the most common techniques is the K-means clustering [26,27]. On the other hand, supervised learning is able to learn based on sample class information to minimize the loss function [28]. Supervised learning was used in the field of cancer classification successfully. Table 1 displays recent development

relevant to ML methods used for cancer classification. Meanwhile, Table 2 depicts the hybrid methods based on supervised and unsupervised learning in cancer classification.

**Table 1.** Supervised Machine Learning Methods.

| Classification Method | Dataset | Description of Ref. | Ref. | Advantages | Limitations |
|---|---|---|---|---|---|
| Support vector machine | ALLAML, DLBCL, Prostate Lung, Lymphoma, MLL, SRBCT, Stjude | Tumor classification using support vector machine and sparse group lasso | [25] | - High accuracy<br>- Prevents overfitting<br>- Flexible kernels selection for nonlinearity<br>- Good ability of generalization | - Complex<br>- Slow training time<br>- Its performance depends on parameter selection |
| | MLL, Lymphoma, Brain, TOX_171, CNS, DLBCL, Lung | Feature selection and tumor classification for microarray data using the relaxed lasso and generalized multi class support vector machine | [28] | | |
| | cDNAs | Classification and validation of cancer tissue samples using microarray expression data | [19] | | |
| | Tumors, Brain Tumors, Leukemia, Lung, SRBCT, DLBCL | Using multicategory support vector machines (MC-SVMs) to the diagnose cancer from gene expression data | [29] | | |
| K-Nearest Neighbors | Colon, Leukemia, Lung, Lymphoma-DLBCL, Ovarian, Prostate | Cancer classification using K-Nearest Neighbors for gene expression data | [30] | - Suitable for multi-modal classes<br>- Independent of the joint distribution of the sample points and their classification | - Lower performance<br>- Depends on the choice of good 'k' value<br>- Performance often varies according to data size |
| | Lung cancer MRI | Cancer detection and classification using an enhanced K-Nearest Neighbors for MRI lung cancer images | [31] | | |
| Naïve Bayes | Wisconsin Breast Cancer dataset (WBCD), lung cancer | Cancer classification using gaussian naive bayes | [32] | - Easy to implement<br>- Requires a small amount of training data<br>- Produce good results most of the cases | - Independence among variables<br>- Less accuracy |
| | Colon cancer | Cancer prediction using naïve bayes model | [33] | | |
| Random Forest | Wisconsin diagnostic breast cancer (WDBC) | Cancer prediction using Random Forest | [33] | - Fast<br>- Scalable<br>- Robust against noise<br>- Does not overfit | - As the number of trees increases, the algorithm slow down |
| | Dermatoscopic images | Skin cancer classification using Random Forest | [34] | | |

**Table 2.** Hybrid of supervised learning and unsupervised learning methods.

| Classification Method | Dimension Reduction Method | Dataset | Description of Ref. | Ref. | Advantages | Limitations |
|---|---|---|---|---|---|---|
| Random Forest | K-means algorithm | Colon cancer, Lung cancer, Prostate tumor | A method of combining feature selection algorithm and classification algorithm using K-means and Random Forest | [35] | - Easy to implement<br>- Scales to large data sets<br>- Adapts easily to new example | - Selecting k manually<br>- Depends on initial values<br>- Clustering data of varying sizes and density |
| Nearest Neighborhood, Support Vector Classifier, Nearest Mean Classifier | | Skin cancer images | Skin cancer classification using K-means algorithm | [36] | | |
| Modified Back Propagation (MBP) | Principal Component Analysis (PCA) | Ovarian, Colon, Leukemia | Cancer detection based on microarray data using PCA and MBP | [37] | - Removes Correlated Features<br>- Efficient performance<br>- Reduces overfitting<br>- Improves visualization | - Independent variables become less interpretable<br>- Requires Data standardization<br>- Loss of information |
| Support vector machine-Recursive Feature Elimination (SVM-RFE) | | Leukemia, Colon, Breast Cancer | Gene selection using PCA for cancer classification | [38] | | |
| k-nearest neighbor (k-NN), artificial neural network (ANN), radial basis function neural network (RBFNN), SVM | Independent Component Analysis (ICA) | Wisconsin diagnostic breast cancer (WDBC) | Feature reduction using ICA for breast cancer detection | [39] | - Efficient computation<br>- High performance for large data sized | - Require more computational for decomposition |
| Kernel SVM (KSVM) | | Brain tumor MRI | Brain tumor detection and classification in MR images using ICA and KSVM | [40] | | |
| Hidden Markov model (HMM) | Scale Invariant Feature Transform | Brain tumor MRI | Brain tumor segmentation and classification based on HMM | [41] | - Strong statistical bases<br>- Efficient performance<br>- Can handle variations in the record structure | - Requires training using annotated data |
| | Singular Value Decomposition (SVD) | Brain tumor MRI | Brain tumor segmentation using HMRF | [42] | | |

### *4.1. Supervised Learning (SL)*

SL is the most common technique for classification problems [43]. The supervised classification algorithms aim at categorizing data from prior information. The class of each test data is determined by joining the features and finding patterns from the training data [3]. Classification involves two phases: (1) a classification algorithm is applied to

the training dataset (2) the model extracted is validated to a test dataset to evaluate the performance of the model and accuracy.

Huo [25] proposed a new method for tumor classification based on the gene's sparse characteristics in gene expression data. The authors have proposed a method that combines both Kruskal–Wallis rank sum test (KW) and sparse group lasso (SGL) for tumor classification. Firstly, this method uses KW for initial selection to remove some redundant genes. Secondly, it uses SGL for further selection to reduce feature genes, and finally, it uses a support vector machine for tumor classification. The author indicated that the proposed method had better performance compared to other methods, such as KNN, Naïve bayes, and Random forest. This proposed method had produced high accuracy with fewer feature genes. Kang [28] proposed a new method for tumor classification via relaxed Lasso and generalized multi-class support vector machine (rL-GenSVM). GenSVM uses regularization parameters to avoid overfitting, reaching high accuracy with fewer feature genes. The optimal parameters for GenSVM are determined by a grid search of 10-fold cross-validation. The results showed that the average accuracy of the GenSVM achieves 4% higher than other classifiers based on the advantages of regularization parameters and radial basis kernel function.

Furthermore, Ayyad [30] proposed a Modified k-nearest neighbor (MKNN) for gene expression cancer classification. Based on KNN, the proposed technique makes use of a new weighting strategy. Six well-known microarray datasets were tested, and the results showed that the classification performance of this technique had increased effectively and time efficiency. Thamilselvan [31] proposed an enhanced K-nearest neighbor (EKNN) for cancer detection and classification in MRI lung cancer images. This proposed method was conducted in 3 stages: the first stage, the morphological method was used in preprocessing to improve the quality of the images, the second stage, the EKNN method was used for identifying cancer, and finally, classifying the images as benign and malignant. The proposed method showed higher accuracy of 97% compared to other methods in image classification. It also produced better results, processing time in 3 s, low misclassification rates, and minimum neighbor distance of 0.20889.

In a study by Kamel [32], the authors proposed a Naïve Bayes algorithm based on Gaussian distribution for cancer classification. The algorithm was tested on two datasets, the Wisconsin Breast Cancer (WBCD) dataset, and the lung cancer dataset. The proposed work used z-score normalization to identify the inefficient value of attributes in the classification that deserved to be zero. The results showed that the proposed work achieved an accuracy of 98% for breast cancer and 90% for lung cancer. Salmi [33] proposed Naïve Bayes model for colon cancer prediction. The authors showed that the proposed model could, therefore, achieve higher classification accuracy and less complexity. In particular, it achieved 95.24% classification accuracy and could, therefore, be an efficient analysis tool.

Octaviani [44] proposed a Random Forest classification for predicting breast cancer data. The proposed method was applied to achieve more accurate and reliable classification performance on cancer microarray data. The data consists of benign and malignant classes. The result showed in this study achieved more than 99% accuracy for the training data. The authors also stated that the proposed method could thus provide more accurate decisions to help the doctors. Nandhini [34] proposed a classification method of skin cancer using Random Forest. The proposed method was applied for classifying skin lesions of seven different types using dermatoscopic images. The result showed that the proposed method achieved 97.3% accuracy on the training dataset.

Very recently, several techniques have been proposed based on supervised learning in cancer classifications [17,45–53]. From the literature, most of the research is focusing on feature selection as well as cancer classification. The Naïve Bayes (NB) classifier has been applied to classified valvular heart disease. The feature selection based on correlation (CFS) was introduced in [45] to select the informative gene of atrial fibrillation (AF) from multi-omics data. The proposed method is accurately classified AF from the valvular heart disease dataset with a precision of 87.5% and AUC of 0.995. Research on identifying biomarkers

that contribute to the disease and cancer are increasing in recent years. Colorectal cancer (CRC) is also the most widely studied in bioinformatics.

Recent work has applied feature selection techniques and machine learning to identify the CRC biomarkers. The Cancer Genome Atlas (TCGA) data are used to perform computational analysis to identify sex-specific biomarkers [46]. On the other hand, several machine learning techniques such as SVM, Random Forest, k-nearest neighbors, and naïve Bayesian tools have been used for the classification and identification of diagnostic markers for major depressive disorder (MDD) [48]. The finding shows that the SVM classifier performed better compared to others in terms of classification accuracy, thus distinguishing MDD samples from healthy, and yielded with an AUC of 0.78. In predicting biomarkers from liver metastasis, several machine learning algorithms were performed, namely logistic regression, Random Forest, SVM, neural network, and CatBoost. Based on the comparative experimental result, the CatBoost algorithm achieved the highest accuracy compared to other algorithms with 99% accuracy. The model was constructed based on 33 informative genes selected from CatBoost algorithm.

scRNA-seq or single-cell RNA sequencing is vital in biomedical research. The work in [50] proposed several classifiers to identify 21 types of cancer and normal tissues using scRNA-seq data. The comparison of machine learning methods was made using NN, kNN, and RF. Based on the result, the NN classifier performed better than other algorithms. Another NN was also introduced to predict biomarkers for disease phenotypes in early sages such as lung cancer [51]. There are many methods and techniques that have been applied using machine learning algorithms for cancer classifications. The classification accuracy is the main concern in the machine learning community. Therefore, work in [52] introduced a procedure for classification using a noisy gene expression dataset. The main contribution was the modified dataset that can improve the accuracy using machine learning algorithms such as SVM, KNN, and Naïve Bayes. Besides cancer classification, other work using supervised machine learning algorithms to detect a DNA copy in cancel cells. The proposed tools called CNAPE are able to predict DNA copy in chromosomes and genes, which produce 80% accuracy.

In summary, many researchers are focusing on supervised machine learning algorithms to identify and predict biomarkers in cancer and disease datasets. SVM classifiers are most widely used and have shown good performance in this direction.

### *4.2. Hybrid of Supervised and Unsupervised Learning (UL)*

Two major UL methods are clustering and principal component analysis (PCA) [22]. Clustering groups subjects with similar characteristics together into clusters [28]. K-means clustering, hierarchical clustering, and Gaussian mixture clustering are the most common clustering algorithms [4]. PCA is commonly used for dimension reduction to make easier and faster computations. PCA projects the data onto a few principal component (PC) directions without losing too much information about the subjects. In certain cases, PCA is first used to reduce data dimension and then used for clustering the subjects into groups.

Aydadenta [35] proposed a method combining feature selection algorithm and classification algorithm using K-means and Random Forest. A clustering K-means algorithm was used to reduce redundancy in microarray data. The features in each cluster were ranked by applying the Relief algorithm. The results showed that for each dataset, namely colon cancer, lung cancer, and prostate tumor, the proposed method achieved 85.87%, 98.9%, and 89% accuracy, respectively. The authors stated that the accuracy of the proposed method was higher than the method without clustering using Random Forest. Mohd [36] proposed a method for classifying two main skin type cancers (melanoma and non-melanoma) using K-means algorithm. In this study, a clustering K-means algorithm was used to segment the skin lesion. The features were extracted from the segmented images using local binary patterns and color percentiles and tested on different classifiers. The results of the proposed method showed good accuracy on different rates of classification.

In a study by Nurfalah [37], the authors proposed a dimension reduction method and classification microarray data using PCA and MBP (modified backpropagation using conjugate gradient). For each dataset, including leukemia, ovarian, and colon cancer, the proposed method yielded 97.14%, 96%, and 76.92% accuracy, respectively. The study showed that PCA and MBP methods combination resulted in a faster training time than the conventional method of backpropagation. Kavitha [38] proposed a gene selection method using PCA and a classification method using SVM-RFE based on cancer microarray data. The research showed that the combination of PCA and SVM-RFE resulted in high accuracy and low error rate compared with the SVM and SVM-RFE algorithms.

Mert [39] proposed a feature reduction method using ICA on tumor classification as benign or malignant. The WDBC dataset dimension was reduced to one feature using ICA. The proposed method was evaluated using several classifiers, such as ANN, k-NN, RBFNN, and SVM. The results showed a slightly decreased accuracy with 30 original features except for RBFNN from 97.53%, 93.14%, and 95.25% to 90.5%, 91.03%, and 90.86%, respectively, while RBFNN increased from 87.17% to 90.49%. The sensitivity rates for the successfully detected malignant samples improved from 93.5% to 96.63% for RBFNN and from 96.07% to 97.47% for SVM, while the others have slightly decreased between 0.96% and 3.09%. This research showed that the proposed method improved the decision support system for diagnostic while reducing computational complexity. In [40], the authors proposed a novel method for the detection and classification of benign and malignant tumors in MR images of the brain. This research used an anisotropic diffusion filter for preprocessing and an active contour model for segmentation. The features were extracted from the tumor MR images using the Daubechies wavelet. The feature vector dimensions were reduced using ICA. A trained SVM with different kernels as KSVM was used for the brain tumor classification. The results showed that the proposed method was effective and fast.

Sharma [41] proposed a method of segmentation and classification based on HMM, which extracted the cancer portion from MRI images of brain cancer. HMM was used to classify abnormal and normal cells based on the properties of images by scanning, segmentation, and classification, and cancer boundary detection. The results showed that the proposed method performed better than the previous method in terms of PSNR, MSE, fault rate dust detection, and accuracy. Mirzaei [42] proposed a method for brain tumor segmentation in MR images using the HMRF classifier, SVD feature extraction method, and wavelet image analysis. The results showed that the proposed method performed better in tumor detection on MR images of the brain.

In summary, these works applied in dimension reduction (feature selection/feature extraction) for cancer classification seem to be promising in terms of the scalability of cancer classification in large-scale models.

## 5. Recent Deep Learning Methods in Cancer Research

Deep learning (DL) algorithms and architectures nowadays attract a lot of attention in the scientific community and research globally. Deep learning is a subset of machine learning algorithms that utilize the advancement of neural networks. It operates by adding multiple hidden layers, the use of activation function, and hyper parameter optimization to process the input and produce the output. With this characteristic, the DL model becomes more complex and more advanced, which gives a lot of benefits to classification tasks. It is more capable of solving complex and large amounts of data compared to the traditional machine learning model. Recently, the application of deep learning has made a significant breakthrough in healthcare, particularly in medical image and cancer classification.

A lot of recent reviews and research are focusing on the deep learning applications applied in cancer diagnosis and prognosis and used genomics dataset [54–60]. Research work in [54] reviewed the application of DL in this field and also summarized its advantages. The authors not only discussed the current literature but also analyzed and recommended ways to advance in this direction. The partitioner or medical specialist is also concerned whether these DL technologies are now matured and ready to be used in

genomics experiments. To address this issue, the work in [55] provided a mini-review of the most distinguished DL model that is already matured in genomics research. The authors also discussed possible challenges and drawbacks and future research directions. The DL is a fast-growing field and accelerates the changes in genomics, especially when involving multimodal data analysis for precision medicine. One of the most used deep learning algorithms is based on the convolution layer. The convolution neural network (CNN) is one of the widely used in image classification. One comprehensive review has been conducted to summarize the usage of machine learning, particularly in DL, in solving medical imaging problems [56]. The results revealed that most of the used imaging datasets are based on MRI, CT, and radiography/mammography. Cancer and disease commonly tacked by DL are neurological and cancer diagnoses. A total of 35% of the research used DL in classification and segmentation. Another research used deep learning for classification the image of the histopathology of canine mammary tumors and also human breast cancer. The author proposed a framework based on VGGNet-16, and the result showed that the accuracy produced by the framework was 97% and 93% for binary classification using the breast cancer and CMT dataset, respectively.

A comparative study using ML and DL algorithms was conducted to analyze the performance of these algorithms in classifying cancer types using microarray gene expression data [57]. The study collected various gene expression datasets of breast, bladder, kidney, lung, and many other diseases and cancer. The comparison was made based on the most widely used algorithm, which is logistic regression and deep learning-based convolution neural network (CNN). The validation of the performance is based on *k*-fold cross-validation. The result shows that CNN is capable of producing 94.43% accuracy compared to traditional machine learning algorithms, with 90.6% accuracy. The interesting finding also shows that the parameter tuning process is not very significant in improving the algorithm accuracy. Two other recent studies also demonstrated that the application of DL for clustering [59] and building a predictive model [58] showed better performance compared to traditional machine learning algorithms, specifically in using multi-omics data for cancer studies.

## 6. Healthcare Dataset for Cancer Classification

In the application of healthcare, AI algorithms need training on the basis of historical data generated from clinical activities, such as diagnosis, screening, and treatment. The historical data is fed to the algorithms to train and learn similar groups and correlations between features [4]. The major sources of health data include physicians notes, diagnostic imaging, and lab test results [5]. These data types have been used by most of the AI techniques in different cases during diagnosis. Specifically, in cancer classification, some of the cancer datasets used by the researchers are Breast Cancer Data Set and Breast Cancer Wisconsin Data Set from UCI Machine Learning Repository [32], some publicly accessible datasets such as microarray data from Kent Ridge Bio-medical Data Set Repository [37], and mini-MIAS database [61].

## 7. Conclusions

Classification problems in the gene expression dataset have largely been studied by researchers in the areas of machine learning and statistics. Recent progress tends to produce robust and advanced methods of classification in order to obtain high accuracy with fewer error rates and with reasonable computation times. Many researchers have proposed methods of cancer classification using various techniques, including traditional ML algorithms based on supervised, unsupervised, and also DL methods that have shown remarkable results. The traditional methods such as SVM and NN perform better compared to others in terms of classification accuracy. Due to the many available methods in this research, the issue of interpretability of the results may arise. With the complexity and high dimensionality of the gene expression data, interpretation of the accuracy from ML and DL methods is not enough. The black-box model such as NN is hard to interpret, especially

when using the gene expression dataset as the input. Many other techniques can be further investigated to study the interpretation of the ML and DL methods for cancer classification, including local and global networks, visualization techniques, and many more. This will open up new possibilities of interpretation studies in cancer classification using ML and DL methods. The huge size of the dataset that is publicly available is also able to accelerate the research efforts. Furthermore, gene expression data based on single-cell RNA sequencing (scRNA-seq) show a promising direction to identify biomarkers that contributes to the cancerous genes. More efforts are still needed to this end, especially when dealing with heterogeneous datasets and multi-class data types. In terms of classification, supervised and DL methods are considered as great interests in this direction. Thus, continued effort is still needed to obtain more robust cancer classification methods in the future.

**Author Contributions:** Conceived the research, A.U.M., Conceptual, M.A.R., N.A.S.; validation, M.S.M., H.W.N.; formal analysis, N.A.S., N.S.N.I., N.B.A.W.; resources A.U.M., M.A.R.; writing-original draft preparation, A.U.M.; writing-review and editing, N.A.S., M.A.R., N.S.N.I.; supervision, M.A.R., N.A.S.; funding acquisition, N.A.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Ministry of Higher Education under Fundamental Research Grant Scheme-RACER, Grant number: RACER/1/2019/ICT02/UMP//3) and Universiti Malaysia Pahang under RDU scheme (Grant number: RDU192624).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data are contained within the article.

**Acknowledgments:** This work was supported by the Ministry of Higher Education under Fundamental Research Grant Scheme-RACER (Grant number: RACER/1/2019/ICT02/UMP//3) and Universiti Malaysia Pahang under RDU scheme (Grant number: RDU192624).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Deepashri, K.S.; Kamath, A. Survey on Techniques of Data Mining and its Applications. *Int. J. Emerg. Res. Manag. Technol.* **2017**, *6*, 198–201.
2. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef]
3. Singh, A.; Thakur, N.; Sharma, A. A review of supervised machine learning algorithms. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; pp. 1310–1315.
4. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [CrossRef] [PubMed]
5. Alloghani, M.; Al-Jumeily, D.; Aljaaf, A.J.; Khalaf, M.; Mustafina, J.; Tan, S.Y. The Application of Artificial Intelligence Technology in Healthcare: A Systematic Review. In *International Conference on Applied Computing to Support Industry: Innovation and Technology*; Springer: Cham, Switzerland, 2020. [CrossRef]
6. Murali, N.; Sivakumaran, N. Review Article Artificial Intelligence in Healthcare—A Review. *Int. J. Modern Comput. Inf. Commun. Technol.* **2018**, *1*, 103–110.
7. Libbrecht, M.W.; Noble, W.S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef]
8. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min.* **2017**, *10*, 1–17. [CrossRef] [PubMed]
9. Goldenberg, S.L.; Nir, G.; Salcudean, S.E. A new era: Artificial intelligence and machine learning in prostate cancer. *Nat. Rev. Urol.* **2019**, *16*, 391–403. [CrossRef]
10. Petegrosso, R.; Li, Z.; Kuang, R. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Brief. Bioinform.* **2019**, *21*, 1209–1223. [CrossRef] [PubMed]
11. Qi, R.; Ma, A.; Ma, Q.; Zou, Q. Clustering and classification methods for single-cell RNA-sequencing data. *Brief. Bioinform.* **2019**, *21*, 1196–1208. [CrossRef]
12. Arora, I.; Tollefsbol, T.O. Computational methods and next-generation sequencing approaches to analyze epigenetics data: Profiling of methods and applications. *Methods* **2021**, *187*, 92–103. [CrossRef] [PubMed]

13. Zielinski, J.M.; Luke, J.J.; Guglietta, S.; Krieg, C. High Throughput Multi-Omics Approaches for Clinical Trial Evaluation and Drug Discovery. *Front. Immunol.* **2021**, *12*, 1–10. [CrossRef]
14. Koteluk, O.; Wartecki, A.; Mazurek, S.; Kołodziejczak, I.; Mackiewicz, A. How do machines learn? Artificial intelligence as a new era in medicine. *J. Pers. Med.* **2021**, *11*, 32. [CrossRef] [PubMed]
15. Avanzo, M.; Trianni, A.; Botta, F.; Talamonti, C.; Stasi, M.; Iori, M. Artificial intelligence and the medical physicist: Welcome to the machine. *Appl. Sci.* **2021**, *11*, 1691. [CrossRef]
16. Yousef, M.; Kumar, A.; Bakir-Gungor, B. Application of biological domain knowledge based feature selection on gene expression data. *Entropy* **2021**, *23*, 2. [CrossRef]
17. Hamzeh, O.; Alkhateeb, A.; Zheng, J.; Kandalam, S.; Rueda, L. Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data. *BMC Bioinform.* **2020**, *21*, 1–10. [CrossRef]
18. Pabby, G.; Kumar, N. A Review on Artificial Intelligence, Challenges Involved & Its Applications. *Int. J. Adv. Res. Comput. Eng. Technol.* **2017**, *6*, 1569–1573.
19. Furey, T.S.; Cristianini, N.; Duffy, N.; Bednarski, D.W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **2000**, *16*, 906–914. [CrossRef]
20. Inza, I.; Larrañaga, P.; Blanco, R.; Cerrolaza, A.J. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.* **2004**, *31*, 91–103. [CrossRef]
21. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef]
22. Dey, A. Machine Learning Algorithms: A Review. *Int. J. Comput. Sci. Inf. Technol.* **2016**, *7*, 1174–1179.
23. Bhola, A.; Tiwari, A.K. Machine Learning Based Approaches for Cancer Classification Using Gene Expression Data. *Mach. Learn. Appl. An Int. J.* **2015**, *2*, 1–12. [CrossRef]
24. Ray, R.; Abdullah, A.A.; Mallick, D.K. Classification of Benign and Malignant Breast Cancer using Supervised Machine Learning Algorithms Based on Image and Numeric Datasets Classification of Benign and Malignant Breast Cancer using Supervised Machine Learning Algorithms Based on Image and Nume. *Int. Conf. Biomed. Eng.* **2019**. [CrossRef]
25. Huo, Y.; Xin, L.; Kang, C.; Wang, M.; Ma, Q.; Yu, B. SGL-SVM: A novel method for tumor classification via support vector machine with sparse group Lasso. *J. Theor. Biol.* **2020**, *486*. [CrossRef] [PubMed]
26. Remli, M.A.; Daud, K.M.; Nies, H.W.; Mohamad, M.S.; Deris, S.; Omatu, S.; Kasim, S.; Sulong, G. K-means clustering with infinite feature selection for classification tasks in gene expression data. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*; Springer: Cham, Switzerland, 2017; Volume 616, pp. 50–57.
27. Sinaga, K.P.; Yang, M.-S. Unsupervised K-Means Clustering Algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [CrossRef]
28. Kang, C.; Huo, Y.; Xin, L.; Tian, B.; Yu, B. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *J. Theor. Biol.* **2019**, *463*, 77–91. [CrossRef]
29. Statnikov, A.; Aliferis, C.F.; Tsamardinos, I.; Hardin, D.; Levy, S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* **2005**, *21*, 631–643. [CrossRef]
30. Ayyad, S.M.; Saleh, A.I.; Labib, L.M. Gene expression cancer classification using modified K-Nearest Neighbors technique. *BioSystems* **2019**, *176*, 41–51. [CrossRef] [PubMed]
31. Thamilselvan, P.; Sathiaseelan, J.G.R. An enhanced k nearest neighbor method to detecting and classifying MRI lung cancer images for large amount data. *Int. J. Appl. Eng. Res.* **2016**, *11*, 4223–4229.
32. Kamel, H.; Abdulah, D.; Al-Tuwaijari, J.M. Cancer Classification Using Gaussian Naive Bayes Algorithm. In Proceedings of the 2019 International Engineering Conference (IEC), Erbil, Iraq, 23–25 June 2019; pp. 165–170. [CrossRef]
33. Salmi, N.; Rustam, Z. Naïve Bayes Classifier Models for Predicting the Colon Cancer. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *546*. [CrossRef]
34. Nandhini, S.; Sofiyan, M.A.; Kumar, S.; Afridi, A. Skin Cancer Classification using Random Forest. *Int. J. Manag. Humanit.* **2019**, *4*, 39–42. [CrossRef]
35. Aydadenta, H.; Adiwijaya, A. A clustering approach for feature selection in microarray data classification using random forest. *J. Inf. Process. Syst.* **2018**, *14*, 1167–1175. [CrossRef]
36. Mohd, A.; Ram, G.K.; Shafeeq, A. Skin cancer classification using K-means clustering. *Int. J. Tech. Res. Appl.* **2017**, *5*, 62–65.
37. Nurfalah, A.; Adiwijaya; Suryani, A.A. Cancer detection based on microarray data classification using PCA and modified back propagation. *Far East J. Electron. Commun.* **2016**, *16*, 269–281. [CrossRef]
38. Kavitha, K.R.; Ram, A.V.; Anandu, S.; Karthik, S.; Kailas, S.; Arjun, N.M. PCA-based gene selection for cancer classification. In Proceedings of the 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, India, 13–15 December 2018. [CrossRef]
39. Mert, A.; Kiliç, N.; Bilgili, E.; Akan, A. Breast cancer detection with reduced feature set. *Comput. Math. Methods Med.* **2015**, *2015*. [CrossRef]
40. Sandhya, G.; Giri, K.; Savitri, S. A novel approach for the detection of tumor in MR images of the brain and its classification via independent component analysis and kernel support vector machine. *Imaging Med.* **2017**, *9*, 33–44.
41. Sharma, S.; Rattan, M. An Improved Segmentation and Classifier Approach Based on HMM for Brain Cancer Detection. *Open Biomed. Eng. J.* **2019**. [CrossRef]

42. Mirzaei, F.; Parishan, M.R.; Faridafshin, M.; Faghihi, R.; Sina, S. Automated Brain Tumor Segmentation in Mr Images Using a Hidden Markov Classifier Framework Trained by Svd-Derived Features. *ICTACT J. Image Video Process.* **2018**, *9*, 1844–1848. [CrossRef]
43. Nasteski, V. An overview of the supervised machine learning methods. *Horizons B* **2017**, *4*, 51–62. [CrossRef]
44. Octaviani, T.L.; Rustam, Z. Random forest for breast cancer prediction. *AIP Conf. Proc.* **2019**, *2168*. [CrossRef]
45. Liu, Y.; Bai, F.; Tang, Z.; Liu, N.; Liu, Q. Integrative transcriptomic, proteomic, and machine learning approach to identifying feature genes of atrial fibrillation using atrial samples from patients with valvular heart disease. *BMC Cardiovasc. Disord.* **2021**, *21*, 1–10. [CrossRef] [PubMed]
46. Hases, L.; Ibrahim, A.; Chen, X.; Liu, Y.; Hartman, J.; Williams, C. The importance of sex in the discovery of colorectal cancer prognostic biomarkers. *Int. J. Mol. Sci.* **2021**, *22*, 1354. [CrossRef] [PubMed]
47. Mitrofanov, A.; Alkhnbashi, O.S.; Shmakov, S.A.; Makarova, K.S.; Koonin, E.V.; Backofen, R. CRISPRidentify: Identification of CRISPR arrays using machine learning approach. *Nucleic Acids Res.* **2021**, *49*. [CrossRef]
48. Zhao, S.; Bao, Z.; Zhao, X.; Xu, M.; Li, M.D.; Yang, Z. Identification of Diagnostic Markers for Major Depressive Disorder Using Machine Learning Methods. *Front. Neurosci.* **2021**, *15*, 1–11. [CrossRef] [PubMed]
49. Shuwen, H.; Xi, Y.; Qing, Z.; Jing, Z.; Wei, W. Predicting biomarkers from classifier for liver metastasis of colorectal adenocarcinomas using machine learning models. *Cancer Med.* **2020**, *9*, 6667–6678. [CrossRef]
50. Kim, B.H.; Yu, K.; Lee, P.C.W. Cancer classification of single-cell gene expression data by neural network. *Bioinformatics* **2020**, *36*, 1360–1366. [CrossRef]
51. Jin, T.; Nguyen, N.D.; Talos, F.; Wang, D. ECMarker: Interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. *Bioinformatics* **2021**, *37*, 1115–1124. [CrossRef] [PubMed]
52. Auwul, R. A Robust Procedure for Machine Learning Algorithms Using Gene Expression Data. *Biointerface Res. Appl. Chem.* **2021**, *12*, 2422–2439.
53. Mu, Q.; Wang, J. CNAPE: A Machine Learning Method for Copy Number Alteration Prediction from Gene Expression. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 306–311. [CrossRef]
54. Huang, S.; Yang, J.; Fong, S.; Zhao, Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Lett.* **2020**, *471*, 61–71. [CrossRef]
55. Koumakis, L. Deep learning models in genomics; are we there yet? *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1466–1473. [CrossRef]
56. Avanzo, M.; Porzio, M.; Lorenzon, L.; Milan, L.; Sghedoni, R.; Russo, G.; Massafra, R.; Fanizzi, A.; Barucci, A.; Ardu, V.; et al. Artificial intelligence applications in medical imaging: A review of the medical physics research in Italy. *Phys. Med.* **2021**, *83*, 221–241. [CrossRef] [PubMed]
57. Tabares-Soto, R.; Orozco-Arias, S.; Romero-Cano, V.; Bucheli, V.S.; Rodríguez-Sotelo, J.L.; Jiménez-Varón, C.F. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Comput. Sci.* **2020**, *2020*, 1–22. [CrossRef] [PubMed]
58. Zhu, W.; Xie, L.; Han, J.; Guo, X. The application of deep learning in cancer prognosis prediction. *Cancers* **2020**, *12*, 603. [CrossRef] [PubMed]
59. Karim, M.R.; Beyan, O.; Zappa, A.; Costa, I.G.; Rebholz-Schuhmann, D.; Cochez, M.; Decker, S. Deep learning-based clustering approaches for bioinformatics. *Brief. Bioinform.* **2021**, *22*, 393–415. [CrossRef] [PubMed]
60. Kumar, A.; Singh, S.K.; Saxena, S.; Lakshmanan, K.; Sangaiah, A.K.; Chauhan, H.; Shrivastava, S.; Singh, R.K. Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. *Inf. Sci.* **2020**, *508*, 405–421. [CrossRef]
61. El Kaitouni, S.E.; Abbad, A.; Tairi, H. A breast tumors segmentation and elimination of pectoral muscle based on hidden markov and region growing. *Multimed. Tools Appl.* **2018**, *77*, 31347–31362. [CrossRef]