applied sciences

MDPI

Brief Report

Pan-Cancer Classification of Gene Expression Data Based on Artificial Neural Network Model

Claudia Cava [1,*], Christian Salvatore [1] and Isabella Castiglioni [2]

[1] Department of Science, Technology and Society, University School for Advanced Studies IUSS Pavia, Palazzo del Broletto, Piazza della Vittoria 15, 27100 Pavia, Italy; christian.salvatore@iusspavia.it
[2] Department of Physics "Giuseppe Occhialini", University of Milan-Bicocca, Piazza dell'Ateneo Nuovo, 20126 Milan, Italy; isabella.castiglioni@unimib.it
* Correspondence: claudia.cava@iusspavia.it

Abstract: Although precision classification is a vital issue for therapy, cancer diagnosis has been shown to have serious constraints. In this paper, we proposed a deep learning model based on gene expression data to perform a pan-cancer classification on 16 cancer types. We used principal component analysis (PCA) to decrease data dimensionality before building a neural network model for pan-cancer prediction. The performance of accuracy was monitored and optimized using the Adam algorithm. We compared the results of the model with a random forest classifier and XGBoost. The results show that the neural network model and random forest achieve high and similar classification performance (neural network mean accuracy: 0.84; random forest mean accuracy: 0.86; XGBoost mean accuracy: 0.90). Thus, we suggest future studies of neural network, random forest and XGBoost models for the detection of cancer in order to identify early treatment approaches to enhance cancer survival.

Keywords: pan-cancer; gene expression; neural network

Citation: Cava, C.; Salvatore, C.; Castiglioni, I. Pan-Cancer Classification of Gene Expression Data Based on Artificial Neural Network Model. *Appl. Sci.* **2023**, *13*, 7355. https://doi.org/10.3390/app13137355

Academic Editor: Lei Zhang

Received: 20 April 2023
Revised: 13 June 2023
Accepted: 19 June 2023
Published: 21 June 2023

1. Introduction

Despite effective advances in research, cancer is one of the leading causes of human deaths, with nearly 10 million deaths in 2020 [1]. Lung cancer and colon and rectum cancer were the most common causes of cancer deaths in 2020, with 1.80 million and 916.000 deaths, respectively [2]. In 2020, 2.26 million new breast cancer cases were diagnosed, making it the most common cancer worldwide [2].

Pan-cancer classification is still a challenge at the molecular level and could be crucial in early cancer diagnosis and for treatment strategies [3]. The challenge for early diagnosis is that the symptoms of many cancers are detected in later stages. The development of a classifier able to identify more cancer types could improve the prognosis of cancer patients, since survival rates dramatically improve in cases of early diagnosis [3].

Gene expression profiles have been associated with different cancers and tissues and have previously been used to build classifiers for different cancer types [4,5]. Differentially expressed genes in many cancer types have been found in genomic regions that play a role in development or carcinogenesis and could influence the expression of downstream genes [6]. Specific patterns of genes have been shown to be significantly altered in many cancers, making gene expression profiles a tool for pan-cancer classification [6]. Gene expression plays a crucial role in the early detection of cancer as it can quantify biochemical processes in tissue and cells [7]. Recently, different genes have been linked to cancer initiation and progression via gene expression analyses [7].

In addition, gene expression has been widely used to identify prognostic and diagnostic gene signatures in cancer and to generate commercial genomic tests. For example, van't Veer et al. [8] proposed a list of 70 prognostic genes in breast cancer and generated a test, MammaPrint, released commercially. Oncotype DX, a qRT PCR-based signature, was developed as the first commercially available test for breast cancer treatment [9].

The enormous quantities of data obtained from high-throughput technologies available in public repositories such as The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) are used for machine learning algorithms, including artificial neural networks [10,11]. However, these datasets present some limitations. They consist of a huge number of gene expression levels or clinical data, as well as noise. Often, the ratio of two classes to be predicted (e.g., normal and cancer) is not balanced, causing biased models with lower performances. The number of genes is usually much greater than the number of samples, requiring a reduction of features dimensionality [12]. As gene expression profiles have high dimensionality, machine learning models often require significant time and resources to train and make a prediction.

Among the wide range of methods applied to reduce data dimensionality and noise, there is principal component analysis (PCA). PCA generates new features, creating a new space from the original feature vectors via a linear transformation [13]. PCA is applied in numerous applications across different fields from social science to biology [14]. Many supervised and unsupervised machine learning algorithms have been applied to gene expression data for cancer prediction [15].

In the last decade, artificial neural networks (ANNs) have been proposed to deal with huge quantities of data. They are a set of algorithms that mimic the human brain to identify complex associations between features and to generate predictive models [16].

Here, we investigated a computational method to classify different cancer types based on ANNs. We compared the performance of the ANN model with the random forest classifier and XGBoost algorithm. XGBoost has been shown to obtain good performance in many application fields including chronic kidney disease and epilepsy diagnosis [17,18]. Several studies have shown that XGBoost obtained better performance in survival prediction in non-small-cell lung cancer and for predicting tissues of origin for 10 different cancer types than other standard machine learning algorithms [19,20]

The results of our study might help in the diagnosis and contribute to planning therapies that could improve cancer survival.

2. Materials and Methods

2.1. Patients, Samples and Gene Expression Data

Gene expression profiles of human tissues and cancer tissues were collected from GEO [21]. The datasets were downloaded with the R software [22] using the GEOquery package [23]. A total of 16 datasets were collected according to the criteria: (i) studies involving cancer/normal tissues, (ii) mRNA expression profiling, (iii) different cancer types.

Table 1 reports the detailed description of considered samples.

Table 1. Cancer type, Gene Expression Omnibus (GEO) ID, size of cancer and normal cases for each dataset. #: number.

Dataset	GEO ID	# of Cancer Samples	# of Normal Samples
Bladder Urothelial Carcinoma	GSE13507	165	10
Breast invasive carcinoma cancer	GSE39004	61	47
Colon adenocarcinoma	GSE41657	25	12
Esophageal carcinoma	GSE20347	17	17
Head and Neck squamous cell carcinoma	GSE6631	22	22
Kidney Chromophobe	GSE15641	6	23
Kidney renal clear cell carcinoma	GSE15641	32	23
Kidney renal papillary cell carcinoma	GSE15641	11	23
Liver hepatocellular carcinoma	GSE45267	48	39

Table 1. *Cont.*

Dataset	GEO ID	# of Cancer Samples	# of Normal Samples
Lung squamous cell carcinoma	GSE33479	14	27
Lung adenocarcinoma	GSE10072	58	49
Prostate adenocarcinoma	GSE6919	65	63
Rectum adenocarcinoma	GSE20842	65	65
Stomach adenocarcinoma	GSE2685	21	8
Thyroid carcinoma	GSE33630	60	45
Uterine Corpus Endometrial Carcinoma	GSE17025	79	12
TOT		749	485

2.2. Data Processing

We standardized each GEO dataset independently, transforming the data distribution per feature to a normal distribution using the function fit_transform in Python. Normalization was performed separately on the training and testing sets.

To avoid unbalanced classes, we applied random oversampling in order to obtain the same number of samples for each class (normal and cancer samples).

Each GEO dataset was divided randomly into two sets: training and testing sets, based on the numbers of cases: 70% of the original dataset for the training and 30% for the testing.

PCA was used to reduce the gene expression data's dimensionality based on 95% of the variance of the training data. We used the same components for the testing dataset [24]. PCA was applied on: (1) the normalized training set, with the PCA parameters saved; (2) the normalized testing set, using the training PCA parameters. PCA was performed using the fit() and transform() functions.

2.3. Neural Network Architecture

We implemented a neural network model consisting of 1 input layer, 2 hidden layers and 1 output layer. The first hidden layer consisted of 17 neurons and the second hidden layer of 8 neurons.

The rectified linear unit (ReLU) was implemented as an activation function at each node of the network [25]. The inputs of the classifier were the key components derived by PCA, while the number of output neurons was the predicted class (cancer or normal). A sigmoid activation function was implemented at the output layer to identify the class to be predicted [26].

We used the Adam optimization algorithm, a modified version of stochastic gradient descent [27]. It was used to assign the parameters that reduce the loss function (binary cross-entropy) as much as possible.

In order to avoid overfitting, we used a "early stopping" function in Keras (https://keras.io/callbacks/#earlystopping accessed on 1 March 2023) that stops the training process when overfitting could be occurring (min_delta = 0.005, patience = 5).

To decrease the time and memory consumption, the model was trained with a batch size = 8 and run for a maximum of 200 epochs.

The training set was used to train the neural network and the testing set to evaluate the accuracy, sensitivity, and specificity of the tested model. The performance of the models was also evaluated using a receiver operating characteristic (ROC) curve, and the area under the curve (AUC).

The training and testing datasets were generated 10 times and the performance measures were summarized as the mean and standard deviation.

The neural network model program was developed in Python using the keras package (version 2.10) [28].

The neural network was compared with a random forest classifier implemented in R with the randomforest package [29].

2.4. Random Forest

Random forest, a supervised learning algorithm, is a decision tree-based model developed by Breiman in 2001 [30].

It consists of an ensemble of trees where each tree considers random samples, and a selection of features is assessed for each node [31].

Each decision tree produces an output of prediction independently. Thus, a final prediction is an average of the different predictions. Considering a vector $x = [x_1, x_2, \ldots, x_n]$ where x represents the model's input features, the final output is defined according to Equation (1):

$$\frac{1}{B}\sum_{b=1}^{B} R_b(x) \quad (1)$$

B is the total number of generated trees and $R_b(x)$ the estimated prediction occurs in the b_{th} tree [32].

Random forest, in our study, was implemented using R package 'randomForest', V4.7-11, (https://cran.r-project.org/web/packages/randomForest/index.html accessed on 1 March 2023). We set the default values of the R package, with 500 trees to grow.

2.5. Extreme Gradient Boosting (XGBoost)

XGBoost, proposed by Chen and Guestrin in 2016, is based on decision trees [33]. It uses gradient tree boosting that makes a prediction via regression trees. It combines the prediction of different regression trees to improve the overall accuracy. Given x, a vector consists of model's input features, the predicted output of XGBoost can be defined according to Equation (2):

$$\sum_{k=1}^{K} f_k(x) \quad (2)$$

where $f_k(x)$ is the output of the k_{th} tree belonging to space of potential regression trees [34,35].

XGBoost was performed using python software with the XGBoost package.

Table 2 shows the applied parameters for the three models. The code is available in: https://github.com/claudiacava/Applied-Sciences accessed on 1 June 2023.

Table 2. Parameters considered for the artificial neural network (ANN), random forest (RF), and XGBoost.

Model	Parameters
ANN	Number of Hidden Layers = 2
	Batch size = 8
	Epochs = 200
	Optimizer = adam
	Losses = binary crossentropy
	Hidden layers activation function = relu
	Output layer activation function = sigmoid
RF	Number of trees = 500
	Minimum size of terminal nodes = 1
	Number of features to be analyzed = (sqrt(p) where p is number of features

Table 2. *Cont.*

Model	Parameters
XGBoost	Loss = mean squared error
	Tree method = gpu hist
	Number of estimators = 100
	Learning rate = 0.3
	Gamma = 0

3. Results

We implemented a pan-cancer classification model using artificial neural networks and gene expression profiles from the GEO database.

Since gene expression profiles contain a high number of genes, we used PCA as a feature selection approach that decreased data dimensionality. This method allows us to find linear combinations of the data that capture the most variance in the data. In addition, to remove over-fitting in the unbalanced dataset, we performed an over-sampling technique.

To evaluate the performance of the classifier, we used a training and testing data set containing gene expression levels of 16 cancer types. To investigate the consistency of our model, we randomly divided the data 10 times into training and testing data sets. We tested the performance of models using the independent testing data that was not included in the training data. We achieved consistent results, considering their average. We compared the performance of the ANN with a random forest classifier and XGBoost. The performance of the methods in terms of sensitivity and specificity are summarized in Table 3.

Table 3. Pan-cancer classification performances of the artificial neural network (ANN), XGBoost and random forest (RF). In bold, we highlighted the best results of the models.

Dataset	Sensitivity			Specificity		
Dataset	**ANN**	**RF**	**XGBoost**	**ANN**	**RF**	**XGBoost**
Bladder Urothelial Carcinoma (GSE13507)	0.85 ± 0.09	0.0 ± 0.0	**0.97 ± 0.03**	0.58 ± 0.22	**1 ± 0.0**	0.79 ± 0.24
Breast invasive carcinoma cancer (GSE39004)	0.80 ± 0.1	0.84 ± 0.09	**0.87 ± 0.06**	**0.82 ± 0.14**	0.79 ± 0.1	0.75 ± 0.07
Colon adenocarcinoma (GSE41657)	0.75 ± 0.24	**1 ± 0**	0.97 ± 0.06	**1 ± 0**	**1 ± 0**	**1 ± 0**
Esophageal carcinoma (GSE20347)	0.85 ± 0.27	**0.97 ± 0.09**	0.89 ± 0.21	**1 ± 0**	0.81 ± 0.28	0.98 ± 0.08
Head and Neck squamous cell carcinoma (GSE6631)	0.69 ± 0.27	**0.96 ± 0.08**	0.93 ± 0.08	**0.92 ± 0.09**	0.84 ± 0.14	0.92 ± 0.11
Kidney Chromophobe (GSE15641)	0.76 ± 0.33	**0.97 ± 0.09**	0.95 ± 0.16	0.84 ± 0.16	0.67 ± 0.45	**0.93 ± 0.14**
Kidney renal clear cell carcinoma (GSE15641)	0.97 ± 0.05	**1 ± 0**	0.91 ± 0.17	**1 ± 0**	**1 ± 0**	**1 ± 0**
Kidney renal papillary cell carcinoma (GSE15641)	0.72 ± 0.37	**1 ± 0**	**1 ± 0**	0.9 ± 0.26	**1 ± 0**	0.93 ± 0.08
Liver hepatocellular carcinoma (GSE45267)	0.85 ± 0.1	**0.97 ± 0.04**	0.92 ± 0.09	0.81 ± 0.12	0.73 ± 0.06	**0.83 ± 0.13**
Lung squamous cell carcinoma (GSE33479)	0.77 ± 0.25	**0.91 ± 0.16**	0.82 ± 0.15	0.87 ± 0.11	0.81 ± 0.16	**0.89 ± 0.20**
Lung adenocarcinoma (GSE10072)	0.79 ± 0.14	**0.99 ± 0.02**	0.94 ± 0.03	0.89 ± 0.08	0.95 ± 0.04	**0.98 ± 0.03**
Prostate adenocarcinoma (GSE6919)	0.57 ± 0.09	0.55 ± 0.16	**0.62 ± 0.17**	0.64 ± 0.14	**0.69 ± 0.17**	0.61 ± 0.17

Table 3. *Cont.*

Dataset	Sensitivity			Specificity		
Rectum adenocarcinoma (GSE20842)	0.93 ± 0.05	**1 ± 0**	0.99 ± 0.02	0.92 ± 0.07	**1 ± 0**	**1 ± 0**
Stomach adenocarcinoma (GSE2685)	0.85 ± 0.11	**0.92 ± 0.19**	0.86 ± 0.2	**1 ± 0**	0.76 ± 0.22	0.96 ± 0.13
Thyroid carcinoma (GSE33630)	0.85 ± 0.1	**0.89 ± 0.07**	0.89 ± 0.1	0.73 ± 0.15	**0.92 ± 0.05**	0.77 ± 0.1
Uterine Corpus Endometrial Carcinoma (GSE17025)	0.94 ± 0.06	0.71 ± 0.24	**0.95 ± 0.07**	0.83 ± 0.17	**1 ± 0**	0.89 ± 0.28
TOT	0.81	0.85	**0.90**	0.86	0.87	**0.89**

The accuracy of three classifiers is shown in Figure 1. For most cancers the accuracy of ANN model era above 0.80. The accuracy was: above 0.85 in colon adenocarcinoma, esophageal carcinoma, kidney renal clear cell carcinoma, rectum adenocarcinoma, stomach adenocarcinoma, and uterine corpus endometrial carcinoma; and below 0.80 in bladder urothelial carcinoma, and prostate adenocarcinoma.

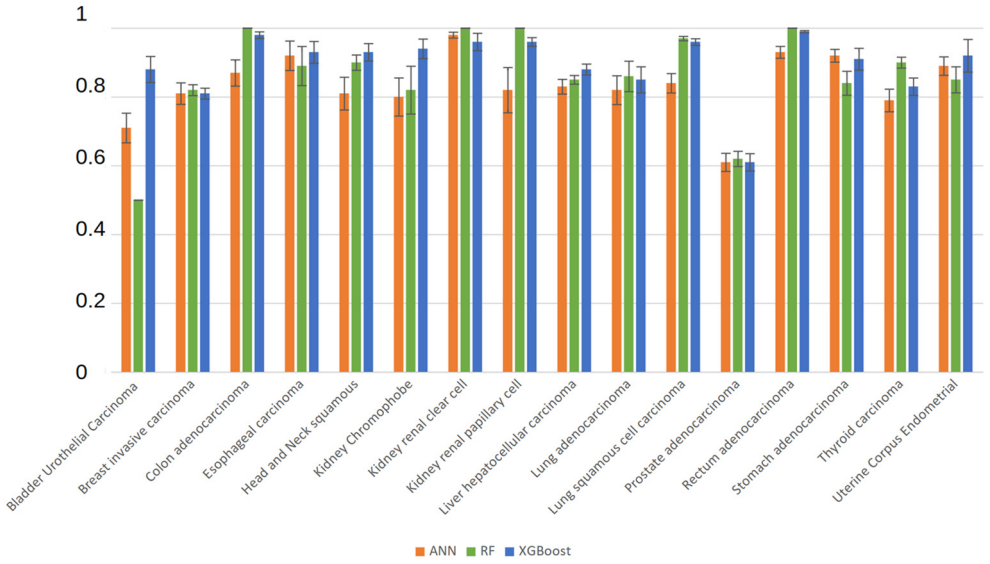

Figure 1. Accuracy with error standard of the classifiers. Artificial neural network (ANN) with blue bars, random forest (RF) with red bars and XGBoost with grey bars.

We compared ANN with random forest classifier and XGBoost. The results of the three classifiers are similar.

T-test demonstrated a statistically significant difference of performance between ANN and XGBoost (*t*-test: sensitivity, *p*-value 0.001, accuracy, *p*-value 0.001, auc, *p*-value 0.049), between ANN and random forest (*t*-test: auc, *p*-value 0.001) and between random forest and XGBoost (*t*-test: auc, *p*-value 0.0003).

The overall accuracy, sensitivity, and specificity of the testing sets for the ANN classifier were 0.83, 0.81 and 0.86, respectively. For the random classifier, we obtained an overall accuracy of 0.86, an overall sensitivity of 0.85 and an overall specificity of 0.87. XGBoost achieved an overall accuracy of 0.90, an overall sensitivity of 0.90 and an overall specificity of 0.89 (Figure 2). The mean AUC values were 0.95 for random forest, 0.92 for XGBoost and 0.89 for ANN.

Figure 3A shows the ROC curves obtained with the ANN model in kidney renal papillary cell carcinoma. The best ROC curves for the random forest classifier were obtained in colon adenocarcinoma, kidney renal papillary cell carcinoma and kidney renal clear

cell and rectum adenocarcinoma (Figure 3B). XGBoost, applied to colon adenocarcinoma, achieved the best ROC curve (Figure 3C).

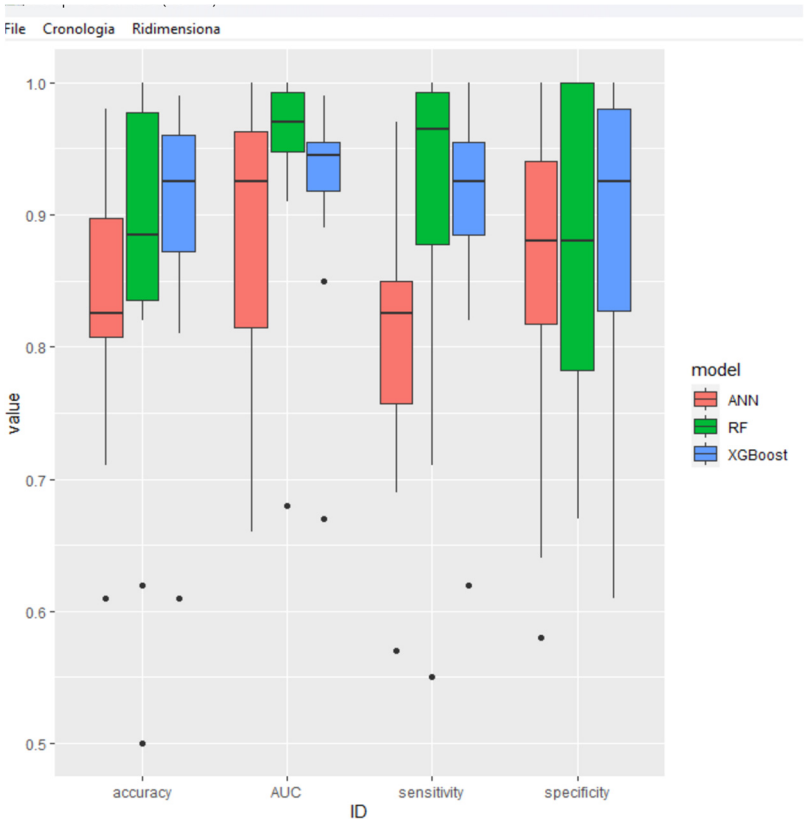

Figure 2. The boxplot shows the accuracy, area under curve (AUC), sensitivity and specificity using the artificial neural network (ANN), random forest (RF), and XGBoost.

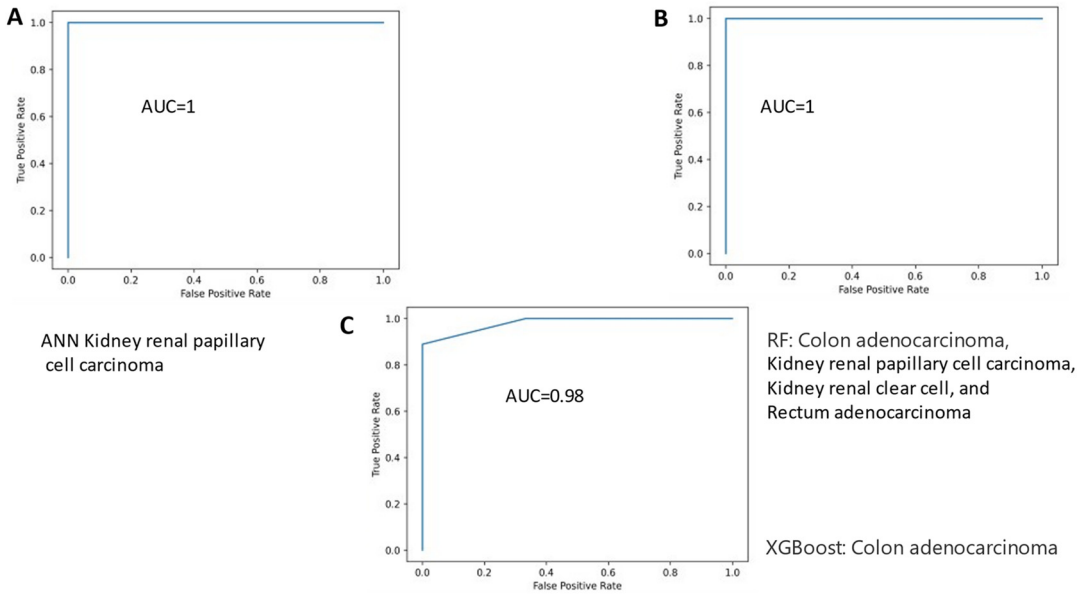

Figure 3. Best ROC curves with: (**A**) the artificial neural network (ANN) model in kidney renal papillary cell carcinoma; (**B**) the random forest (RF) model in colon adenocarcinoma, kidney renal papillary cell carcinoma and kidney renal clear cell and rectum adenocarcinoma; (**C**) the XGBoost model in colon adenocarcinoma.

4. Discussion

In this study, we used the gene expression profiles of 749 cancer patients and 485 normal samples of 16 different cancer types from GEO database to provide a pan-cancer analysis using a deep learning approach compared with random forest classifier and XGBoost models.

Previous studies have already shown a higher performance of ANNs compared with logistic regression in cancer [36,37].

However, in pan-cancer analysis, there is no clear evidence of the predictive accuracy of ANNs in cancer diagnosis (tumor vs. normal prediction). There are few studies that have demonstrated the abilities of ANNs to predict cancer samples, and the applications to date have been mainly performed to few cancer types [38,39]. In addition, we applied a novel recent algorithm, XGBoost, which obtained good classifier performance in cancer in other studies [17,18]. Many studies have demonstrated the role of XGBoost in the prediction of origin of tissues in cancer, and few studies in tumor vs. normal prediction [17,18].

In our study, the performance of the ANN model is compared with a random forest classifier and XGBoost to carry out a comparative analysis of classification models. The classification was performed on gene expression levels of cancer and normal tissues, focusing on two main aspects in the ANN model: first, we used PCA as method to reduce data dimensionality; second, we propose the use of the Adam optimization algorithm.

The ANN model achieved accuracies greater than 0.80 in most cancers based on key components derived from PCA. Random forest achieved high performances, close to the ANN model. We conclude that the ANN model and random forest have similar high performances. Random forest obtained a slightly better performance than the ANN, but it was not significant (t-test: p-value > 0.5). XGBoost achieved a better performance in accuracy (p-value = 0.001), sensitivity (p-value = 0.001) and AUC (p-value = 0.049) compared with the ANN. No statistically significant difference was revealed between the ANN and random forest in sensitivity, specificity and accuracy.

These findings are consistent with reports presented in other studies. Indeed, recent studies have suggested neural networks as promising tools in classification analysis using gene expression data [3]. Ainscough et al. demonstrated that random forest and deep learning approaches obtained high and similar performance using cancer sequencing data [40].

In summary, our results demonstrated that our ANN classifier obtained classification performance similar to random forest using a limited number of samples.

The models are only tested on one dataset for each cancer type. This study found high and similar performance of the three models, but further studies should be performed on other datasets in future works, and the associations between gene expression levels and genetic aberrations should be also investigated.

Despite the interesting results of our study, there are some limitations to be noted. First, the ANN is difficult to configure, as the model is dependent on the structure of the network used, the choice of activation functions, the regularization approach used, the depth of the network and other factors still. Second, the ANN model should be further applied to larger samples, as several studies reported that the performance of a classifier based on an ANN increases with the number of samples [41]. Another limitation of the current approach is that we modelled and predicted the model considering a training and testing dataset but not a validation set. This is due to small size of the pan-cancer dataset, which limits its usage only to testing evaluation. The lack of a validation set could induce an overfitting bias that needs to be explored in a future work. The authors will expand their research in order to increase the models' accuracy using other validation sets and multi-omics data.

5. Conclusions

We presented an analysis that compared some of the commonly used machine learning approaches. We applied the methods to 16 different cancer types and compared the results.

Although the three classifiers achieved high and similar performance (neural network mean accuracy: 0.84; random forest mean accuracy: 0.86; XGBoost mean accuracy: 0.90), we found that XGBoost obtained the better performance. In the testing set, XGBoost also showed the highest performance in sensitivity and specificity (0.90 and 0.89, respectively). In terms of AUC values, random forest obtained the best prediction results (0.95). However, we suggest deepening the three models given the high and similar performance obtained.

The good performance of the models demonstrated the efficiency of the classifiers based on gene expression levels, and we suggest that these models could be extended to other phenotypes and integrated in future studies.

In addition, subsequent analyses should be addressed: (i) to identify a gene signature, (ii) to use our models for single gene expression analysis, (iii) to achieve a standardization of data collection, and normalization.

Author Contributions: Conceptualization, C.C.; methodology, C.C.; formal analysis, C.C. writing—original draft preparation, C.C. and C.S.; writing—review and editing, C.C. and C.S.; supervision, I.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly archived datasets analyzed are reported in https://www.ncbi.nlm.nih.gov/geo/ accessed on 1 March 2023.

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

References

1. Ferlay, J.; Ervik, M.; Lam, F.; Colombet, M.; Mery, L.; Piñeros, M.; Znaor, A.; Soerjomataram, I.; Bray, F. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer. 2020. Available online: https://gco.iarc.fr/today (accessed on 1 February 2021).
2. World Health Organization. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 1 February 2023).
3. Gore, S.; Azad, R.K. CancerNet: A unified deep learning network for pan-cancer diagnostics. *BMC Bioinform.* **2022**, *23*, 229. [CrossRef] [PubMed]
4. Cava, C.; Castiglioni, I. In Silico perturbation of drug targets in pan-cancer analysis combining multiple networks and pathways. *Gene* **2019**, *698*, 100–106. [CrossRef]
5. Cava, C.; Bertoli, G.; Castiglioni, I. Portrait of Tissue-Specific Coexpression Networks of Noncoding RNAs (miRNA and Lncrna) and mRNAs in Normal Tissues. *Comput. Math. Methods Med.* **2019**, *2019*, 9029351. [CrossRef] [PubMed]
6. Cava, C.; Bertoli, G.; Castiglioni, I. In silico identification of drug target pathways in breast cancer subtypes using pathway cross-talk inhibition. *J. Transl. Med.* **2018**, *16*, 154. [CrossRef] [PubMed]
7. Alharbi, F.; Vakanski, A. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. *Bioengineering* **2023**, *10*, 173. [CrossRef]
8. van't Veer, L.J.; Dai, H.; van de Vijver, M.J.; He, Y.D.; Hart, A.A.M.; Mao, M.; Peterse, H.L.; van der Kooy, K.; Marton, M.J.; Witteveen, A.T.; et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **2002**, *415*, 530–536. [CrossRef]
9. Paik, S.; Shak, S.; Tang, G.; Kim, C.; Baker, J.; Cronin, M.; Baehner, F.L.; Walker, M.G.; Watson, D.; Park, T.; et al. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med.* **2004**, *351*, 2817–2826. [CrossRef]
10. Wang, Y.; Xu, X.; Maglic, D.; Dill, M.T.; Mojumdar, K.; Ng, P.K.-S.; Jeong, K.J.; Tsang, Y.H.; Moreno, D.; Bhavana, V.H.; et al. Comprehensive Molecular Characterization of the Hippo Signaling Pathway in Cancer. *Cell Rep.* **2018**, *25*, 1304–1317.e5. [CrossRef]
11. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for functional genomics data sets—Update. *Nucleic Acids Res.* **2013**, *41*, D991–D995. [CrossRef]
12. Guyon, I. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
13. Makiewicz, A.; Ratajczak, W. Principal Components Analysis (PCA). *Comput. Geosci.* **1993**, *19*, 303–342. [CrossRef]
14. Świetlicka, I.; Kuniszyk-Jóźkowiak, W.; Świetlicki, M. Artificial Neural Networks Combined with the Principal Component Analysis for Non-Fluent Speech Recognition. *Sensors* **2022**, *22*, 321. [CrossRef] [PubMed]
15. 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**, *6*, e270. [CrossRef] [PubMed]

16. Michie, D.; Spiegelhalter, D.J.; Taylor, C.C. Machine Learning, Neural and Statistical Classification. *Technometrics* **1994**, *37*, 459. [CrossRef]
17. Ogunleye, A.A.; Wang, Q.-G. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *17*, 2131–2140. [CrossRef]
18. Torlay, L.; Perrone-Bertolotti, M.; Thomas, E.; Baciu, M. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Inform.* **2017**, *4*, 159–169. [CrossRef]
19. Huang, Z.; Hu, C.; Chi, C.; Jiang, Z.; Tong, Y.; Zhao, C. An Artificial Intelligence Model for Predicting 1-Year Survival of Bone Metastases in Non-Small-Cell Lung Cancer Patients Based on XGBoost Algorithm. *BioMed Res. Int.* **2020**, *2020*, 3462363. [CrossRef]
20. Zhang, Y.; Feng, T.; Wang, S.; Dong, R.; Yang, J.; Su, J.; Wang, B. A Novel XGBoost Method to Identify Cancer Tissue-of-Origin Based on Copy Number Variations. *Front. Genet.* **2020**, *11*, 585029. [CrossRef]
21. Gene Expression Omnibus. Available online: http://www.ncbi.nlm.nih.gov/geo (accessed on 1 January 2023).
22. R Development Core Team. *Computational Many-Particle Physics*; Springer: Berlin/Heidelberg, Germany, 2008.
23. Davis, S.; Meltzer, P.S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **2007**, *23*, 1846–1847. [CrossRef]
24. Moolayil, J. Tuning and Deploying Deep Neural Networks. In *Learn Keras for Deep Neural Networks*; Apress: Berkeley, CA, USA, 2019. [CrossRef]
25. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, PMLR, Fort Lauderdale, FL, USA, 11–13 April 2011; Gordon, G., Dunson, D., Dudík, M., Eds.; Mir Press: Banksmeadow, Australia, 2011; Volume 15, pp. 315–323.
26. Yuan, Y.; Bar-Joseph, Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 27151–27158. [CrossRef]
27. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
28. Chollet, F. Keras, GitHub. 2015. Available online: https://github.com/fchollet/keras (accessed on 1 March 2023).
29. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22. Available online: https://CRAN.R-project.org/doc/Rnews/ (accessed on 1 March 2023).
30. Fatahi, R.; Nasiri, H.; Homafar, A.; Khosravi, R.; Siavoshi, H.; Chelgani, S.C. Modeling operational cement rotary kiln variables with explainable artificial intelligence methods—A "conscious lab" development. *Part. Sci. Technol.* **2022**, *41*, 715–724. [CrossRef]
31. Walker, A.M.; Cliff, A.; Romero, J.; Shah, M.B.; Jones, P.; Gazolla, J.G.F.M.; Jacobson, D.A.; Kainer, D. Evaluating the performance of random forest and iterative random forest based methods when applied to gene expression data. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 3372–3386. [CrossRef]
32. Homafar, A.; Nasiri, H.; Chelgani, S. Modeling coking coal indexes by SHAP-XGBoost: Explainable artificial intelligence method. *Fuel Commun.* **2022**, *13*, 100078. [CrossRef]
33. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
34. Chelgani, S.C.; Nasiri, H.; Tohry, A.; Heidari, H. Modeling industrial hydrocyclone operational variables by SHAP-CatBoost—A "conscious lab" approach. *Powder Technol.* **2023**, *420*, 118416. [CrossRef]
35. Amjad, M.; Ahmad, I.; Ahmad, M.; Wróblewski, P.; Kamiński, P.; Amjad, U. Prediction of Pile Bearing Capacity Using XGBoost Algorithm: Modeling and Performance Evaluation. *Appl. Sci.* **2022**, *12*, 2126. [CrossRef]
36. Hanai, T.; Yatabe, Y.; Nakayama, Y.; Takahashi, T.; Honda, H.; Mitsudomi, T.; Kobayashi, T. Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. *Cancer Sci.* **2003**, *94*, 473–477. [CrossRef]
37. Pergialiotis, V.; Pouliakis, A.; Parthenis, C.; Damaskou, V.; Chrelias, C.; Papantoniou, N.; Panayiotides, I. The utility of artificial neural networks and classification and regression trees for the prediction of endometrial cancer in postmenopausal women. *Public Health* **2018**, *164*, 1–6. [CrossRef]
38. Lee, K.; Jeong, H.-O.; Lee, S.; Jeong, W.-K. CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network. *Sci. Rep.* **2019**, *9*, 16927. [CrossRef]
39. Yuan, Y.; Shi, Y.; Li, C.; Kim, J.; Cai, W.; Han, Z.; Feng, D.D. DeepGene: An advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinform.* **2016**, *17*, 243–256. [CrossRef] [PubMed]
40. Ainscough, B.J.; Barnell, E.K.; Ronning, P.; Campbell, K.M.; Wagner, A.H.; Fehniger, T.A.; Dunn, G.P.; Uppaluri, R.; Govindan, R.; Rohan, T.E.; et al. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet.* **2018**, *50*, 1735–1743. [CrossRef] [PubMed]
41. Alwosheel, A.; van Cranenburgh, S.; Chorus, C.G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J. Choice Model.* **2018**, *28*, 167–182. [CrossRef]