Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 185 (2021) 264–273

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference Theme: Big Data, IoT, and AI for a Smarter Future
Malvern, Pennsylvania, June 16-18, 2021

Comparative Study of Disease Classification Using Multiple Machine Learning Models Based on Landmark and Non-Landmark Gene Expression Data

Xiaoqin Huang [a#], Jian Sun [b#],Satish Mahadevan Srinivasan [a]*,Raghvinder S Sangwan [a]

[a] *Pennsylvania State University, Engineering Department, 30 Swedesford Rd, Malvern, Pennsylvania, 19355, USA*
[b] *German Center for Neurodegenerative Diseases (DZNE), Otfried-Müller-Str. 23, Tübingen, 72076, Germany*

Abstract

This study compares disease classification based on landmark and non-landmark gene expression data, and clinical variable using multiple machine-learning models. The influence of the number of principal components and the genes were also investigated. The results indicate that the ANN model has the best accuracy for disease type prediction among all the models, model using 95 principal components has better accuracy than that of 25 principal components, and the greater number of genes used, the higher the prediction accuracy. Models using landmark genes demonstrated better accuracy than the models using non-landmark genes especially with 95 PCs across all the models except for the decision trees. The optimal model was one that uses landmark genes with 95 PCs as features for an ANN classifier. The AUC measures obtained on the test set were 0.98,0.98,1 and 0.96 for *Autoimmune*, *Bacteremia*, *Cancer* and *Healthy* classes respectively, and the accuracy for the respective classes were 97.56%, 95.65%, 95.65%, and 58.82%. The ANN model demonstrated a good capability of distinguishing between the true positives and the false positives, and it resulted in high prediction accuracy for the 3 disease classes (*Autoimmune*, *Bacteremia*, *Cancer*), but it misclassified some instances from the *Healthy* class to the *Autoimmune* and *Bacteremia* class, likely due to a wide range of gene expression level for the *Healthy* class.

Keywords: Landmark Gene; Disease Classification; Machine Learning; Artificial Neural Network; Gene Expression Analysis

* Corresponding author. Tel.: 6104279288; fax:
E-mail address: sus64@psu.edu

1877-0509
10.1016/j.procs.2021.05.028

1. Introduction

Gene expression profiling has been extensively used to capture the gene expression patterns in cellular responses to disease and drug treatment. The change of gene expression level can lead to phenotype variation. Capturing the pattern of gene expression for specific disease type is vital since it can provide insight for disease diagnosis as well as treatment [1-4].

Due to the high cost of whole genome expression profiles and high correlations in gene expression, researchers have selected a special set of about 978 landmark genes, which has been reported to be highly reproducible and is suitable for computational inference [5]. Chen *et. al.* [6] have used deep learning techniques to infer non-landmark gene expression based on landmark gene expression and they have reported that the deep learning techniques outperformed linear regression models by 15.33%. McDermott *et. al.* [7] used the L1000 LINCS dataset, and one privately produced gene expression dataset, to classify cell types and subtypes using *K* Nearest Neighbour (KNN) classifiers, Decision Trees (DT), Random Forests (RF), linear classifiers, and two neural classifiers: Feed-Forward Artificial Neural Networks (FF-ANN) and Graph Convolutional Neural Networks (GCNN). They reported that GCNNs have better performance when dealing with large number of samples. Li *et. al.* [8] have used XGBoost to predict gene expression values for non-landmark genes based on landmark gene and have reported that their method outperforms the existing models for determining the gene expression values. Carly *et. al.* [9] have reported that 978 landmark genes have better clustering capabilities of microarray and clinical datasets compared to any randomly chosen 978 genes from the whole genome.

Machine learning has been applied to a variety of analysis related to gene expression and disease study [10]. Xiao *et. al.* have used different models to predict cancer based on RNA-seq data from diverse datasets and have reported that deep learning-based multi-model ensemble methods have better performance on all the datasets [11]. Liu *et. al.* has applied different machine learning techniques including Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and RF to develop predictive models that can distinguish between TCMR and STA samples based on RNA-seq data and clinical variables. They have reported that the RF classifier achieved the best specificity and sensitivity [12]. Tabares-Soto *et. al.* [13] have applied both machine learning and deep learning models to classify cancer type based on microarray gene expression data and have reported that convolutional neural networks resulted in better accuracies *i.e.* ~ 94%. Alanni *et. al.* [14] have proposed a gene selection programming (GSP) method using SVM for classifying cancer type based on microarray datasets. The gene set selected by their GSP has shown superior performances in cancer classification compared to the gene sets selected by other gene selection methods.

In this work, we aim to investigate the significance of landmark genes for disease type prediction. Multiple classification models, including DT, RF, SVM, ANN and eXtreme Gradient Boosting (XGBoost) were explored in this study to compare the disease type classification based on a combination of clinical variables and landmark /non-landmark genes. Our objective here is to identify an optimal model for disease type prediction based on the gene expression data and the clinical variables (gender and disease type). Our preliminary studies have demonstrated that the landmark genes have the capabilities to better distinguish between different tissue types using unsupervised learning (Figure S1). However, it would be very beneficial to see if the landmark genes can be used as features to predict different disease types with better accuracies.

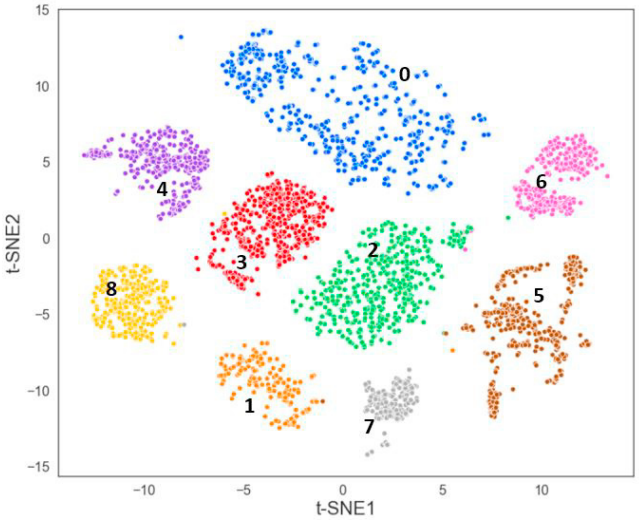

Fig. S1. t-SNE plot for tissue clustering based on landmark gene expression
(0 - Human aortic endothelial cells, 1 - Peripheral blood mononuclear cells, 2 - Breast tumor, 3 - Pre-treatment bone marrow, 4 - Lymphoblastoid cell line, 5 - Leukemic cells, 6 - Colon cancer tissue, 7 - Kidney, 8 - CD138 purified myeloma plasma cells)

2. Method and Materials

2.1 Classifiers

This study explores several different classifiers including DT, RF, SVM, XGBoost and ANN. DT is one of the successful approaches applied in the bioinformatics field for both regression and classification. It is represented by a tree-like structure where inner nodes represent a test on a feature, each branch represents outcome of a test, and each leaf node represents class label, and the decision is made after traversing the corresponding branch of the tree and reaching its leaf . Based on the approaches of measuring information to choose the best feature as the root node, there are two algorithms, one that uses the entropy as the measurement, the other using the Gini impurity [15]. In this study, we chose Gini impurity.

RF has also been used for both classification and regression. It is one of the widely used algorithms due to its simplicity. It consists of multiple decision trees. In RF, the dataset is partitioned into smaller subsets and for each subset RF builds its own decision tree with leaf nodes and decision nodes. The results from all the subset trees are averaged to provide an output of the model with small fluctuations. This algorithm is suitable for handling large datasets and tries to utilize all the features in the dataset. It also helps to reduce overfitting.

ANN is a computational structure inspired by the study of the biological neural processing. There are several different types of ANN ranging from simple to very complex. ANN represents a highly parallelized dynamic system with a directed graph topology. The ANN acronym comprises a variety of feed forward networks that are commonly called back-propagation networks. The back propagation refers to the method for computing the error gradient for a feed forward network. A general topology of the ANN consists of several input neurons that receive a normalized numerical input from each of the variables in the dataset. These normalized values are then multiplied by factors, known as connection weights. The net product of these multiplications is summed up which then becomes the net input that enter into an activation function that calculates the outputs of the hidden neurons [16]. In this study the ANN model was constructed with one hidden layer of 10 neurons, “relu” was used as the activation function, and Adam was used as the optimizer with a learning rate of 0.001.

The SVM algorithm tries to find a hyperplane in a *high-dimensional space obtained by a non-linear transformation from the original n-dimensional* . The hyperplane is used to classify the data points. Vectors that

approach the hyperplane are known as support vectors. Removing the support vectors will lead to alteration in the position of hyperplane. We have used a gaussian kernel to implement SVM in this study.

XGBoost is an ensemble learning method and is based on gradient boosting. Gradient boosting works on the principle of boosting an ensemble of weak learners, typically DTs, iteratively by shifting focus towards problematic observations that were difficult to predict in previous iterations. It builds the model in a stage-wise fashion like other boosting methods do, but it generalizes them by allowing optimization of an arbitrary differentiable loss function [17]. Loss function depends on the type of problem being solved. In this case of classification problems, logarithmic loss is used. In boosting, at each stage, unexplained loss from prior iterations would be optimized rather than starting from scratch. Trees are added one at a time and existing trees in the model are not changed. The gradient descent procedure is used to minimize the loss when adding trees. We have employed a grid based hyperparameter tuning for the XGBoost model to identify an optimal model for the dataset used in this study.

Principle component analysis (PCA) is a classic dimension reduction approach, which has been extensively used in gene expression studies. It seeks linear combinations of original features, termed principal components (PCs), that can effectively represent effects of all the original features. PCs are orthogonal to each other and choosing the number of PCs depends on the variance that can be explained by the PCs for the original features [18]. The number of PCs chosen is much smaller than the number of original features. In this study, we use linear PCA and compare the disease classification with different number of PCs.

As the target classes are imbalanced, we have opted techniques to balance the dataset. In this study we have performed over sampling [19] of the minority classes in order to match the number of instances in the minority class with that of the number of instances in the majority class.

2.2 Dataset

The dataset used in this study, previously analysed by Chen *et. al.* 2016 [6], consists of 978 landmark genes within a set of 22,678 genes. These genes were measured across 129,157 observations/samples in the dataset. These observations were associated with metadata, including clinical variables, such as gender and disease type which were queried from the NCBI-GEO database using the methodology documented in [5, 20]. We encoded the disease type variables into four classes namely, *Autoimmune*, *Bacteriamia*, *Cancer* and *Healthy*, based on specific disease types specified in the original dataset.

A subset dataset consisting of landmark gene expressions was constructed and was combined with the clinical variables (LM978). For comparison purposes, 20 different groups of 978 non-landmark were randomly selected from the L1000 dataset and were combined with the clinical variables (Non-LM978). We also constructed five different groups of 300 and 500 genes randomly selected from the 978 landmark genes and 21,700 non-landmark genes including the clinical variables (dataset LM300, LM500 Non-LM300, Non-LM500.

Table S1. Descriptions about the datasets used

Dataset Name	Size (Row*Column)	Variables
LM978	828*980	978 landmark genes+2 clinic variables
Non-LM978	828*980	978 non-landmark genes+2 clinic variables
LM300	828*302	300 landmark genes+2 clinic variables
Non-LM300	828*302	300 non-landmark genes+2 clinic variables
LM500	828*502	500 landmark genes+2 clinic variables
Non-LM500	828*502	500 non-landmark genes+2 clinic variables

2.3 Performance Measures

In this study, we have tried to address a multi-class classification problem. We report the results of our experiments using *Accuracy*, *Recall*, *Precision* and *AUC*. *Accuracy* is the ratio of the number of truly predicted samples to the total number of samples in the test dataset. *Precision* is the fraction of relevant instances among the

retrieved instances and *Recall* is the fraction of the relevant instances that have been retrieved over the total number of relevant instances. We provide the receiver operating characteristic curve (ROC) that is commonly used to visualize the performance of a classifier and compute the Area Under the Curve (AUC). The ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters, namely the True Positive Rate (TPR) and False Positive Rate (FPR). AUC ranges in value from 0 to 1. Generally, the classifier with high AUC or high recall is highly desirable for any classification problem.

2.4 Classification of disease types using multiple predictive models

We performed different predictive modeling techniques on the datasets namely the LM978, Non-LM978, LM300, LM500, Non-LM300, Non-LM500. In all these datasets, the variable *disease type* was considered as the response variable and the remaining variables were used as predictors. Each dataset was split into training and test dataset as 80:20. The training dataset was used to train the model and the test dataset was used for model evaluation. In order to account for the imbalance within the training dataset we performed oversampling with respect to the target variable.

The training dataset was oversampled and PCA was performed separately on training and test set. The objective here was to reduce the dimension of the datasets and choose a subset of principle components (25PCs and 95PCs) across each dataset. We then applied different machine learning models including the DT, RF, XGBoost, SVM and ANN on each dataset with different number of principal components as the predictors and the disease type as the response variable. All the models were evaluated using the classification accuracy as the performance measure. The accuracy was obtained by averaging the accuracies of different datasets. The optimal model was selected based on the classification accuracy, and other performance metrics including the ROC curve, Precision, Recall and AUC.

3. Results and discussion

3.1 Clustering

Figure 1 demonstrates the clustering result after dimensionality reduction using the original distribution of disease types in a two-dimensional space. On the left-hand side, four groups were separated well by *k*-means clustering since the *k*-means clustering aims at minimizing the Euclidian distance between the points and the centroid. On the right-hand side of the figure, the true distribution of disease types does not seem to be separated well since there are many overlaps among the disease types *Bacterimia*, *Autoimmune* and *Healthy*. However, the *Cancer* disease type was well separated from the others. This observation can be attributed to the reduction of dimensions using which it is difficult to reflect the real distributions of the data points. In addition to that the disease types cannot be separated by a simple linear method that is merely based on the distance of the datapoints. Therefore, it is evident that there is a need for advanced machine learning techniques to distinctly classify the different disease types.

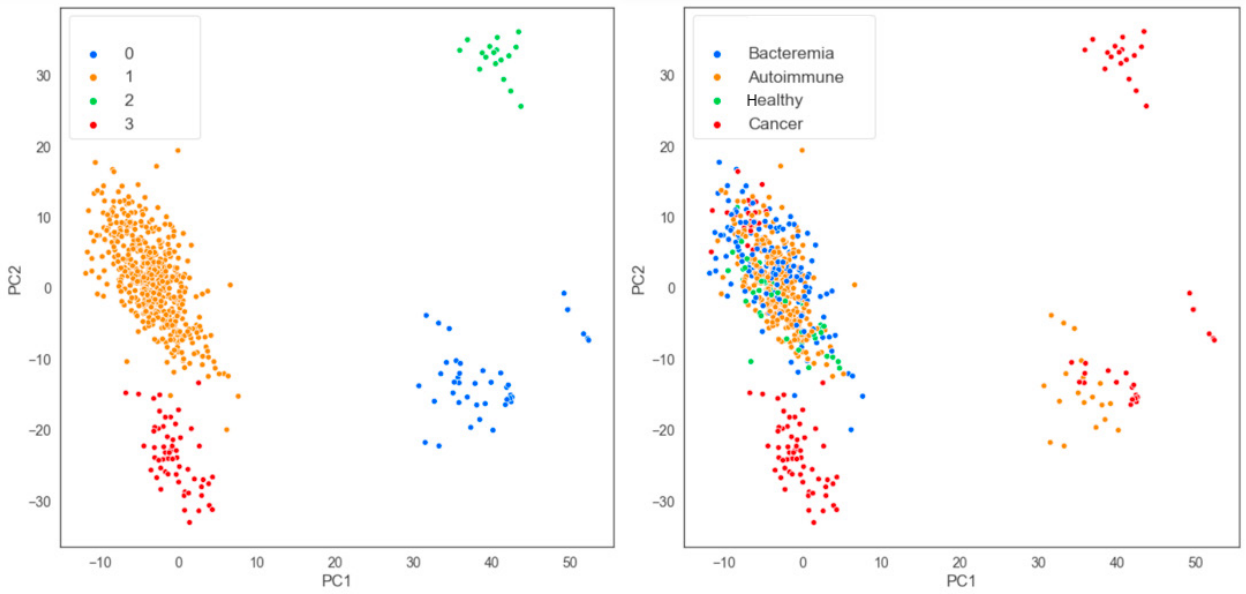

Fig. 1. Distribution of 4 clusters based on landmark gene after PCA (left: k-means clustering result; right: true label of disease type)

3.2 Classification

3.2.1 Comparing predictive models based on landmark gene, non-landmark gene and principal components as predictors

The accuracies of the different predictive models for disease type classification is shown in Table 1. The ANN classifier recorded the highest accuracy for disease type classification using both the 25 PCs and the 95 PCs as the predictors. However, we noted a significant increase in the prediction accuracy for the ANN classifier using landmark genes as predictors (95 PCs) compared to using the non-landmark genes as predictors (95 PCs). Following ANN, SVM recorded the second-best accuracies when using both the landmark and non-landmark genes as predictors when considering the first 95 PCs. However, when considering only the first 25 PCs in case of both the landmark and non-landmark genes the XGBoost classifier outperformed the SVM classifier. Irrespective of 25 PCs or 95 PCs, landmark or non-landmark genes as predictors, the DT recorded the least performance in classifying the disease type. 95 PCs has higher accuracy in all the models except for DT, which can be attributed to the ratio of explained variance; 25 PCs explained only about 75% of the variance while 95 PCs explained a total of 90% of the variance. The lower accuracy for DT with 95 PCs can be attributed to the fact that it exhibits overfitting in a higher dimension dataset [21].

Table 1. Comparison of accuracy based on different models under 25 and 95 PCs

	25 PCs				95 PCs			
Model	Landmark	CI_lm	Non-landmark	CI_non_lm	Landmark	CI_lm	Non-landmark	CI_non_lm
DT	0.821	(0.809,0.833)	0.793	(0.737,0.848)	0.720	(0.715,0.725	0.773	(0.730,0.816)
RF	0.839	(0.815,0.863)	0.860	(0.831,0.889)	0.869	(0.859,0.879)	0.843	(0.813,0.872)
XGboost	0.869	(0.852,0.886)	0.872	(0.842,0.903)	0.899	(0.871,0.927	0.873	(0.847,0.898)
SVM	0.827	(0.791,0.863)	0.846	(0.802,0.890)	0.905	(0.899,0.911)	0.882	(0.862,0.902)
ANN	0.899	(0.877,0.924)	0.885	(0.850,0.919)	0.929	(0.909,0.949)	0.896	(0.868,0.924)

CI_lm: confidence interval of landmark gene; CI_non_lm: confidence interval of non-landmark gene

RF and XGBoost are both ensemble methods based on DT. The RF builds many trees based on a subset of features and makes predictions based on the majority voting. On the other hand, XGBoost builds many tree models and the subsequent model learns from the mistakes of the previous model. Therefore, these models have better performance than the DT model. SVM has been reported as a powerful method in many biomedical tasks [10], but in this dataset, SVM only has better performance in higher dimensional data (95PCs.

Overall with 95 PCs, landmark genes have better prediction accuracy than the non-landmark genes, but with only 25 PCs, there is no apparent difference indicating that in higher dimensions landmark genes have a higher representativeness than non-landmark genes.

3.2.2 Influence of the number of genes on the model performance

Table 2 summarizes the results of the prediction accuracy for different models. Here we varied the number of genes used as the predictors for classifying the disease types. On each dataset we performed PCA and selected the top 95 PCs. Overall, the ANN classifier has the best accuracy when we choose 978 and 300 landmark genes to classify the disease types. However, the SVM classifier demonstrated a significant increase in the prediction accuracy when we choose 500 landmark and non-landmark genes to predict the disease types. Overall, it is evident that the more the number of genes the higher the prediction accuracies of the classifier in case of both the landmark and non-landmark genes. While using only 300 or 500 genes do not have much difference in case of landmark genes, we see a significant difference in the improvement of the prediction accuracies with 500 non-landmark genes rather than using 300 non-landmark genes. This observation supports the fact that the number of genes is important for disease type classification particularly in case of the non-landmark genes since more information is included within the additional genes. Based on Table 2, we can clearly conclude that the 978 landmark genes have the potential to boost the performance of the ANN classifier in classifying the different disease types. Except for DT, majority of the models have shown better prediction accuracies for disease classification using landmark genes as predictors as oppose to using the non-landmark genes as predictors.

Table 2. Comparison of accuracy and confidence interval based on landmark and non-landmark gene with different gene number under 95 PCs

Model	lm978	CI_lm978	lm500	CI_lm500	lm300	CI_lm300	non-lm978	CI_non_lm978	non-lm500	CI_non_lm500	non-lm300	CI_non_lm300
DT	0.720	(0.715,0.725)	0.793	(0.726,0.860)	0.790	(0.784,0.796)	0.773	(0.730,0.816)	0.744	(0.655,0.833)	0.750	(0.678,0.821)
RF	0.869	(0.859,0.879)	0.854	(0.830,0.877)	0.854	(0.842,0.866)	0.843	(0.813,0.872)	0.836	(0.804,0.868)	0.810	(0.775,0.844
XGboost	0.899	(0.871,0.927)	0.886	(0.849,0.923	0.876	(0.871,0.882)	0.873	(0.847,0.898)	0.850	(0.812,0.888)	0.839	(0.810,0.868)
SVM	0.905	(0.899,0.911)	0.904	(0.876,0.930)	0.876	(0.847,0.906)	0.882	(0.862,0.902)	0.876	(0.845,0.907)	0.860	(0.822,0.897)
ANN	0.929	(0.909,0.949)	0.896	(0.851,0.942)	0.900	(0.889,0.912)	0.896	(0.868,0.924)	0.871	0.843,0.900)	0.833	(0.786,0.880)

CI_lm: confidence interval of landmark gene; CI_non_lm: confidence interval of non-landmark gene

3.2.3 ANN Model evaluation using multiple evaluation metrics

Based on the results of the comparative models (see Table 1 and Table 2), it is evident that the ANN classifier using landmark genes (LM978, using the first 95 PCs) has the best accuracy, therefore, we further evaluate this classifier using multiple evaluation metrics. Figure 2 shows the accuracy curve across different epochs. The accuracy of the ANN model increased from 0 to ~1 after ~20 epochs for both the training and the validation (a subset of the training dataset) dataset. The model fits well across the different epochs since the accuracy of both the training and validation is almost overlapping, so there is no signal of overfitting.

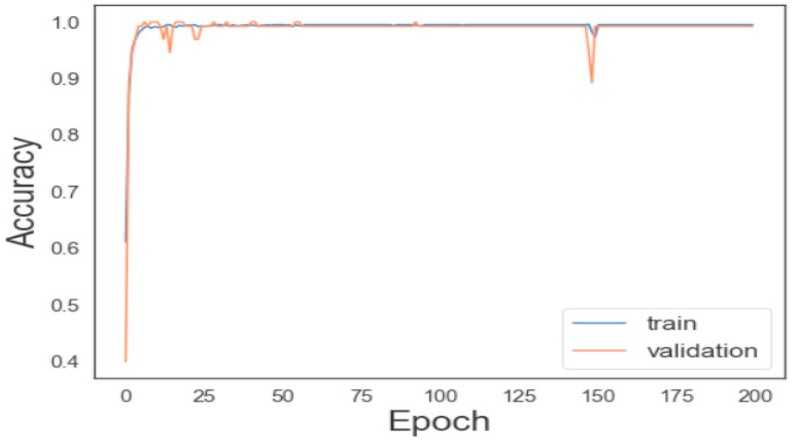

Fig. 2. Accuracy of ANN model under 95PCs for landmark genes

Figure 3 shows the ROC curve for the ANN model on the testing dataset (LM 978). The AUC for the disease classes *Autoimmune*, *Bacteremia*, *Cancer* and *Healthy* are 0.98,0.98,1.00, and 0.96 respectively, indicating that the model has a better potentiality to distinguish the true positives from the false positives across each disease classes. Figure 4 shows the confusion matrix on the testing (LM 978) dataset. The prediction accuracies for the disease classes *Autoimmune*, *Bacteremia*, *Cancer* and *Healthy* are 97.56%, 95.65%, 95.65%, and 58.82% respectively, among which the disease types *Autoimmune*, *Bacteremia* and the *Cancer* have higher accuracies, while for the *Healthy* class, the accuracy is low; 23.53% and 17.65% of the instances belonging to the *Healthy* class has been misclassified into *Autoimmune* and *Bacteremia* classes respectively. Misclassifications in the *Healthy* class can be attributed to the fact that there might be a greater variation of gene expression in healthy class when compared to other disease classes [22-23].

Upon taking a closer look at the results, none of the *Healthy* class instances were misclassified into the *Cancer* class (and also the other way around) indicating that the ANN model can clearly distinguish the instances between these two classes. This observation has clinical significance as there is a strong indication of no false positives as well as no false negatives.

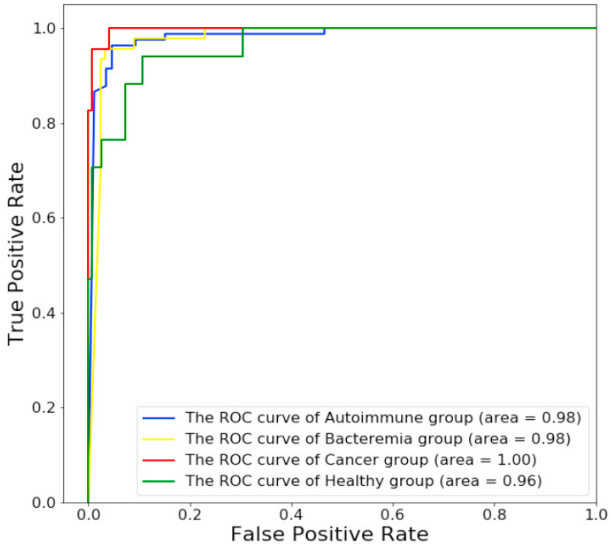

Fig. 3. ROC curve of disease prediction based on ANN model under 95 PCs for landmark genes

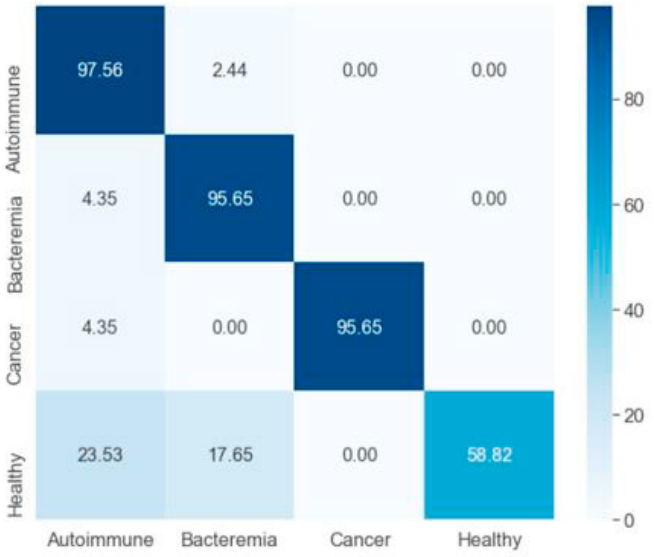

Fig. 4. Confusion matrix (values are in percentage) for the ANN model on the testing dataset (LM978)

Table 3 records the precision and recall of the ANN classifier on the LM978 test dataset using the top 95PCs (landmark genes). All the four classes have high precision, well above 90%. Recall for the *Autoimmune*, *Bacteremia*, and *Cancer* disease classes ranges from 96-98% indicating that majority of the instances in these classes are correctly labelled. However, for the *Healthy* class the recall is only 0.59, which means that ~41% of the actual *Healthy* instances were incorrectly labelled as *Autoimmune* and *Bacteremia* (see Table 3 and Figure 4). This might be improved by acquiring more clinic variables for the model training in the future study since healthy class is closely related to other variables, such as age, race, etc, or by setting weights to each class to deal with the imbalance instead of oversampling.

Table 3. ANN Model evaluation metrics using 95 PCs (landmark genes)

	Precision	Recall
Autoimmune	0.92	0.98
Bacteremia	0.90	0.96
Cancer	1.00	0.96
Healthy	1.00	0.59

4. Conclusion

Based on the multiple (25/95 PCs, 978/500/300 genes) comparative study of disease classification using both landmark and non-landmark genes, and multiple machine learning models (DT, RF, XGBoost, SVM, ANN), the results indicate that the ANN model has the best accuracy among all the models. Also, it is evident that models exhibit better accuracies using 95 PCs rather than using 25 PCs. In this study, we have also demonstrated that the more the number of genes used as predictors for disease classification, the higher the prediction accuracies across most of the classifiers. Models (except for DT) with landmark genes as predictors have better accuracies than with the non-landmark genes especially when we consider the first 95 PCs across all the models.

The limitation of this study is that not enough clinical variables were included in the dataset due to the limited number of samples available after cleaning the clinical variables and merging them with the gene expression datasets. We chose the variables with the maximum number of samples obtained after cleaning and merging with the

gene expression data, so only information related to the gender and disease type were included. Future studies can work on collecting more related clinical variables from various data sources to improve the classification model.

References

[1] Celis, J. E., Kruhøffer, M., Gromova, I., Frederiksenb, C., østergaarda, M., Thykjaerb, T., et al. (2000) "Gene expression profiling: monitoring transcription and transition products using DNA microarrays and proteomics." *FEBS Lett.*, **480 (1)**, 2-16.

[2] Mclachlan, G. J., Do, K-A, and Ambroise, C. (2005) *Microarrays in Gene Expression Studies[M]//Analyzing Microarray Gene Expression Data*. John Wiley & Sons, Inc.

[3] Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., et al. (2006). "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." *Lancet,* **17 (2)**, 154-155.

[4] Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2016) "Prediction and validation of disease genes using HeteSim Scores." *IEEE/ACM Trans. Comput. Biol. Bioinf.,* **14 (3)**, 687-695.

[5] Subramanian, Aravind, Rajiv Narayan, Steven M. Corsello, David D. Peck,et al. (2017). "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles." *Cell,* **171(6)**: 1437-1452.e17.

[6] Chen Y., Li Y., Narayan R., et al. (2016) "Gene expression inference with deep learning. Bioinformatics," **32(12)**: 1832-1839.

[7] McDermott M.B.A., Wang J., Zhao W, et al. (2020) "Deep Learning Benchmarks on L1000 Gene Expression Data." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. DOI 10.1109/TCBB.2019.2910061,

[8] Li W, Yin Y, Quan X and Zhang H. (2019) "Gene Expression Value Prediction Based on XGBoost Algorithm." *Front. Genet.*, 10: 1077.

[9] Clayman C.L., Srinivasan S.M., and Sangwan R.S. (2020) "K-means clustering and principal components analysis of microarray data of L1000 Landmark Genes." *Procedia Computer Science*,168: 97-104.

[10] Libbrecht M.W., and Noble W.S. (2015) "Machine learning in genetics and genomics." *Nat Rev Genet,* **16(6)**: 321-332.

[11] Xiao Y., Wu J., Lin Z., et al. (2018) "A deep learning-based multi-model ensemble method for cancer prediction." *Computer Methods and Programs in Biomedicine*, 153: 1-9.

[12] Liu P., Tseng G., Wang Z., et al. (2019) "Diagnosis of T-cell–mediated kidney ejection in formalin-fixed, paraffin-embedded tissues using RNA-Seq–based machine learning algorithms." *Human Pathology*, 84, 283-290.

[13] Tabares-Soto R., Orozco-Arias S., Romero-Cano V., et al. (2020) "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data." *PeerJ Comput. Sci*., 6: e270. DOI 10.7717/peerj-cs.270.

[14] Alanni R., Hou J., Azzawi H., et al. (2019) "A novel gene selection algorithm for cancer classification using microarray datasets." *BMC Medical Genomics*, 12:10.

[15] Che D., Liu Q., Rasheed K., Tao X. (2011). "Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics." Arabnia H., Tran QN. (eds) *Software Tools and Algorithms for Biological Systems. Advances in Experimental Medicine and Biology*, vol 696, Springer, New York, NY. https://doi.org/10.1007/978-1-4419-7046-6_19

[16] Gharehchopogh, F. S., and Mohammadi, P. (2013) "A Case Study of Parkinson's disease Diagnosis using Artificial Neural Networks*." Internation Journal of Computer Applications*, **73(19)**.

[17] Chen, T., and Guestrin, C. (2016) "XGBoost: A Scalable Tree Boosting System." Proceedings of the *KDD'*, 16, 1-10.

[18] Ma S., and Dai Y. (2011) "Principal component analysis based methods in bioinformatics studies." Briefings In Bioinformatics, **12(6)**: 714-722.

[19] Sáez J.A., Krawczyk B., Woźniak M. (2016) "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets." *Pattern Recognition*, 57, 164-178.

[20] Enache et al. (2019) "The GCTx format and cmap {Py, R, M, J} packages: resources for optimized storage and integrated traversal of annotated dense matrices." *Bioinformatics*, **35 (8)**: 1427-1429.

[21] Kotsiantis S. B. (2013) "Decision trees: a recent overview." *Artif Intell Rev*, 39:261-283.

[22] Buscher, K., Ehinger, E., Gupta, P., et al. (2017) "Natural variation of macrophage activation as disease-relevant phenotype predictive of inflammation and cancer survival." *Nat Commun,* **8,** 16041.

[23] Chalancon, G., Ravarani, C. N., *et al*. (2012) "Interplay between gene expression noise and regulatory network architecture." *Trends in genetics: TIG*, **28(5)**, 221-232.