IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 April 2023, accepted 15 May 2023, date of publication 25 May 2023, date of current version 14 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3280066

RESEARCH ARTICLE

Exploration of the Relevance of MicroRNA Signatures for Cancer Detection and Multiclass Cancer Classification

MATTHEW ACS, RICHARD ACS, CHARLES BRIANDI, EYAN EUBANKS, ONEEB REHMAN, AND HANQI ZHUANG (Senior Member, IEEE)
Department of Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA
Corresponding author: Oneeb Rehman (orehman@fau.edu)

ABSTRACT miRNA expression profiles are heterogeneously expressed among cancer types, with miRNAs serving as highly tissue specific tumor suppressors and oncogenes. Machine learning methodologies have been used to develop high performance pan-cancer classification models and identify potentially novel miRNA biomarkers for clinical investigation. However, it is important to understand how such data science techniques correlate to established biological processes to advance integration into clinical environments. This research aims to assess how the top miRNA features selected by machine learning models relate to clinically and biologically verified miRNA biomarkers. We developed Support Vector Machine and Random Forest machine learning models for cancer classification, iteratively adding cancer classes to the multiclass models. The relationship between the selected top features (miRNAs) and clinically verified miRNA biomarkers was assessed through percent relevance, i.e., the number of verified miRNAs vs the number of selected features. We found that as the number of cancer classes increased, the performance metrics decreased, yet the percentage relevance of the miRNA feature selection signature slightly increased before stabilizing. Additionally, after conducting principal component analysis, the non-cancer tissues from all samples had very similar expression visualizations, while all cancerous tissues had unique profiles. The results indicated that models with a greater number of cancer classes shift towards focusing on cancer-diverse miRNAs of greater relevance with characterized functionality. This work suggests that miRNAs may be highly unique to specific cancerous tissues and can be strong biomarkers for detection and classification, but current verified biomarkers fall toward more cancer-wide miRNAs when detecting cancer.

INDEX TERMS Cancer classification, cancer detection, miRNA expression, principal component analysis (PCA), random forest, support vector machine.

I. INTRODUCTION

MicroRNAs, or miRNAs, are a class of small non-coding RNAs that play an important role in regulating gene expression [1]. miRNAs achieve gene regulation by targeting specific messenger RNAs and marking them for degradation or translational repression [1]. Previous studies have shown that miRNAs play an important role in the development of numerous human pathologies. Ha reviewed the role of miRNAs in cardiovascular disease, stating that miRNAs are important for regulating cardiomyocyte self-renewal and differentiation, as well as for normal cardiac structural integrity [2]. Chen et al. associated miRNAs with nervous system functions, highlighting that a deficiency of miR-133b can be observed in the midbrain of Parkinson's patients [3]. They further found that various high-profile diseases, such as diabetes, Alzheimer's disease, cardiac hypertrophy, acquired immune deficiency, and numerous cancers are closely associated with miRNA functionality [3].

Cancer, in particular, is one of the most devastating diseases in the United States, accounting for more than $209.9 billion in total care costs in 2005 and surpassing

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

heart disease as the leading killer of Americans under the age of 85 [4]. The promising association of miRNAs with cancer development has fueled significant research interest in understanding any potentially causal links. Numerous studies have documented the role of miRNAs in controlling biological functions such as cellular proliferation, differentiation, and apoptosis [3], which form the basis of cancer development. Moreover, Esquela-Kerscher and Slack have shown that miRNA expression profiling can serve as a more accurate method of classifying cancer subtypes than protein-coding gene expression profiling [5]. Notably, such classification is enabled by the heterogenous miRNA expression patterns among cancer types, with miRNAs serving as highly tissue specific tumor suppressors and oncogenes [6]. This has led to an effort to categorize and understand the miRNAs within cancer expression profiles that most significantly contribute to cancer development. Fu et al. summarized breast cancer miRNA biomarkers and the functional pathways that had been identified [7]. Various other studies have similarly summarized potential miRNA biomarkers for cancers such as lung, kidney, liver, and colon among others [8], [9], [10], [11]. Unsurprisingly, this has spurred interest in investigating miRNA expression profiling data as a potentially less invasive diagnostic tool for early detection. With the development of large publicly available miRNA datasets such as the NCI Genomic Data Commons (GDC) Data Portal [12] and the NCBI Gene Expression Omnibus [13], the challenge of miRNA diagnostics has been approached from a data analytics standpoint.

Recently, numerous machine learning approaches have been developed for classifying various cancers based on miRNA expression profiles. Kalecky et al. used the Cancer Genome Atlas dataset to distinguish between basal-like 1 and basal-like 2 triple negative breast cancers [14]. Yang et al. created a 16-miRNA signature based diagnostic model for lung adenocarcinoma [15]. Various studies have also developed successful classifiers for esophageal squamous cell carcinoma, stage-2 colon cancer, and kidney cancer subtypes among others [16], [17], [18]. Besides specific single-cancer studies, the emergence of large miRNA databases such as the Cancer Genome Atlas have enabled studies to investigate the feasibility of multiclass cancer classification. Telonis et al. utilized binarized isomiR profiles to distinguish between 32 TCGA cancers with an average sensitivity of 90% [19]. Lopez-Rincon et al. utilized a 100-miRNA signature to classify a dataset of 8023 samples of 28 different types of cancer [20]. They aimed to identify reliable miRNA signatures that can be used for clinically relevant prediction tasks. Other studies attempted to classify smaller subsets of cancers, ranging from 21 TCGA cancers utilizing a support vector machine in one study [21], to 20 different tumor anatomical sites using a deep neural network in another [22]. These previous studies demonstrate the feasibility of miRNAs to be used as biomarkers for cancer classification, as well as the importance of further exploring the role of miRNA signatures in improving clinical classification applications.

This research aims to understand potential factors involved in creating a multiclass cancer classifier based on miRNA expression data. To better understand the implications of creating a multiclass miRNA cancer diagnostic tool, we utilized Support Vector Machine (SVM) and Random Forest models to distinguish between an increasingly large subset of cancerous tissues. For each iteration of our methodology, we added another cancerous tissue class to the models, utilizing a single class for all non-cancer tissues. The 20-feature miRNA signature was established using feature extraction techniques, and the accuracy, precision and recall from model cross validation is reported for each iteration. The purpose of this study is threefold; to understand the change in success metrics as more cancer types are introduced to the subset, to understand how the 20-miRNA signature changes as more cancer types are introduced to the subset, and to understand the characteristics of the full dataset via principal component analysis. Unlike previous studies which have only focused on miRNA feature signatures for a final multiclass dataset, this study tracks changes in clinical and biological relevance after each addition of a cancerous tissue type. This provides insights into potential relationships between the overall clinical relevance of the feature extraction signature and the success metrics of the models. Additionally, this study analyzes the feasibility of using a multi-tissue miRNA cancer signature as a generalizable signature for single class cancer detection in a number of prominent cancers.

II. MATERIALS AND METHODS

Our methodology utilized an iterative process that applied several key techniques to a continually increasing dataset of miRNA expression quantification data. The techniques that were used are described in the subsections below and the procedure used is described in the final subsection.

A. DATASET

The data that we used for this project was obtained from the GDC data portal [12]. We selected all miRNA expression quantification data from TCGA projects that had at least 100 samples to ensure that sufficient data is available to model each class. The data was separated into classes based on the primary sample site. Overall, 20 data classes were used from the GDC data portal, including samples originating from the breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts, cervix uteri, soft tissue, retroperitoneal and peritoneum, esophagus, adrenal gland, pancreas, and testis. Table 1 shows the number of cancerous and non-cancerous samples for each primary sample site. Each data sample originally had four features associated with each of 1881 miRNAs. These features were miRNA_ID, read_count, reads_per_million_miRNA_mapped, and cross-mapped. We removed all

TABLE 1. Distribution of samples across classes.

Cancer Primary Sample Site	Number of Cancerous Samples	Number of Non-cancerous Samples	Total Number of Samples
Breast	1103	104	1207
Kidney	545	71	616
Corpus Uteri	546	22	568
Thyroid Gland	514	59	573
Bronchus and Lung	521	46	567
Prostate Gland	499	52	551
Brain	530	0	530
Ovary	499	0	499
Stomach	446	45	491
Colon	450	8	458
Skin	450	2	452
Bladder	418	19	437
Liver and Intrahepatic Bile Ducts	375	50	425
Cervix Uteri	309	3	312
Soft Tissue	118	0	118
Retroperitoneal and Peritoneum	101	0	101
Esophagus	185	13	198
Adrenal Gland	151	3	154
Pancreas	179	4	183
Testis	156	0	156

features except the reads_per_million_miRNA_mapped and miRNA_ID so that the models could use the concentration of miRNA for each miRNA to create the classification model. Finally, the data was encoded so that 0 signified the joint non-cancer class and integers starting at 1 signified the respective cancer class for each primary sample site.

B. FEATURE SELECTION

Feature selection was performed to identify the miRNAs that were the most influential in separating the classes. This is important because it allows us to compare how the most relevant miRNAs change as more primary sample sites are added to the classification model. The feature selection method that we used was based on mutual information, which measures the dependency between two random variables. The mutual information for two continuous random variables is described by (1), where the marginal density of X is described by (2) and the marginal density of Y is described by (3) [23]. In order to perform feature selection, the mutual information between the class labels and data points was maximized using a subset of the total features. The 20 best features were selected and used for comparison with biologically and

clinically relevant miRNAs as described in subsection G of materials and methods.

$$I(X,Y) = \iint dxdy\mu(x,y)\, log\frac{\mu(x,y)}{\mu_x(x)\mu_y(y)} \quad (1)$$

$$\mu_x(x) = \int dy\mu(x,y) \quad (2)$$

$$\mu_y(y) = \int dx\mu(x,y) \quad (3)$$

C. DATA PREPROCESSING

Data preprocessing was performed before proceeding to feature selection or model training. First, a transformation specified by (4) was applied to each data point. Next, the data was standardized to a mean of zero and a unit variance, and the parameters used to standardize the data were obtained from the training data and applied to both the training and test datasets.

$$log_2(x+1) \quad (4)$$

D. CLASSIFICATION MODELS

A SVM model and a random forest model were used for this purpose. These models are classical supervised machine learning approaches that have been successfully used to classify miRNA expression profiles in a number of studies [19], [21]. A SVM seeks to separate two or more classes using hyperplanes to create decision boundaries that maximize the width of the gap between the classes. The SVM that we used was configured using a linear kernel since we did not find any significant improvement with other kernels. A Random Forest model uses the aggregate decision of several decision trees to determine the classification of the datapoint. We decided to use 200 decision trees for each Random Forest model based on our experimentation that showed a high level of performance using 200 trees.

E. K-FOLD CROSS VALIDATION

A K-fold cross validation procedure was used to evaluate each model for each iteration of the study. We utilized 5-fold cross validation, in which the data is shuffled, and 5 subsets are created. Then, 80% (4 subsets) of the data samples are used for training and the remaining 20% (1 subset) for testing, and this process is repeated five times by changing the subsets that are used for testing and training. This ensures that all samples are tested at least once, and it mitigates any inconsistencies that may occur in a classical 80-20 train test split due to imbalances in our relatively small dataset [24]. Cross validation was also used in accordance with steps outlined in the Data Analysis Protocol which is part of the US-FDA MAQC-II initiative that aims to establish best practices for reproducibility across different technologies and laboratories and evaluate the utility of these technologies in clinical and safety assessments [25]. The accuracy (5), precision (6), and recall (7) performance metrics are reported based on the cross validation results using the following measures: true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

$$(TP+TN)/(TP+TN+FP+FN) \quad (5)$$

$$(TP)/(TP+FN) \quad (6)$$

$$(TP)/(TP+FP) \quad (7)$$

F. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis, or PCA, is a type of dimensionality reduction technique that was used to visualize the miRNA data and has been previously shown to effectively reduce redundant information in miRNA expression data [26]. The miRNA expression profiles in our dataset have 1881 different features, therefore visualization utilizing all features would be infeasible. Thus, PCA attempts to embed the features distributed in the higher dimensional space into a lower dimensional space that can be visualized. In our case, we conducted and graphed a two dimensional and three dimensional PCA analysis, allowing us to visualize the space occupied by each class in two and three dimensions.

G. RETRIEVAL OF CLINICALLY RELEVANT MIRNAS

A literature review was performed to create a list of identified miRNA biomarkers for cancer. First, a review was conducted using the PubMed database to find miRNA biomarkers that have robustly studied biological roles and well-established associations with cancer. The resulting miRNA biomarkers were organized into a table for further analysis with our miRNA feature selection signatures to establish the percentage relevance. For the purposes of our study, we define clinical (or biological) percentage relevance as the fraction of miRNAs identified in the 20-miRNA feature selection signatures that are present in the list of clinically (or biologically) verified biomarkers from literature. We performed another literature review using PubMed to create a comprehensive list of all biomarkers identified through established biological, non-data science, methodologies such as northern blotting, qRT-PCR, digital PCR (dPCR), microarrays, and next generation sequencing (NGS) techniques [27]. This list was then used to establish the percentage relevance for each dataset's feature extraction in our iterative methodology.

H. PROCEDURE

In this study, we iteratively added classes to a classification model. We started by only adding two classes, non-cancer and breast cancer. The data was then preprocessed according to the preprocessing outlined in subsection C of materials and methods. The mutual information of top-20 feature selection was executed on the dataset and the relevant miRNAs were recorded. An SVM and Random Forest with the parameters outlined in subsection D of materials and methods were trained separately using the 5-fold cross validation procedure. The accuracy, precision, and recall of each model were recorded. The top-20 features were also compared with miRNA biomarkers identified in the relevant literature and the percentage relevance described in subsection G of materials and methods was calculated. This process

TABLE 2. Best-characterized cancer-associated microRNAs [28], and [29].

Clinically Verified miRNA	Functionality	Biological Role
Let-7 family	Tumor-suppressor	Direct regulator of some important oncogenes, cell cycle and cell proliferation genes, apoptosis, and RNase III nuclease [33]
miR-15/-16	Tumor-suppressor	Inhibit cell proliferation, tumorigenicity and induce apoptosis [34]
miR-29 family	Tumor-suppressor	Regulation of cell proliferation, differentiation, and apoptosis [35]
miR-34 family	Tumor-suppressor	Repressing tumor progression by involving in epithelial-mesenchymal transition [36]
miR-26a	Tumor-suppressor	Validated target genes are involved in cell metabolism, proliferation, differentiation, apoptosis, invasion and metastasis [37]
miR-200 family	Tumor-suppressor	Cancer initiation and metastasis through targeting transcription factors [27]
miR-155	Oncogene	Important role in hematopoiesis, targets regulatory proteins for myelopoiesis and leukemogenesis [38]
miR-21	Oncogene	Important roles in immune cell development and function, Epithelial-to-Mesenchymal Transition and Fibrosis, oncogene regulation, and early development [39]
miR-221/-222	Oncogene	Regulate functions of cancer cells to proliferate, differentiate, and invade [40]
miR-17/92	Oncogene	Cell cycle, proliferation, apoptosis and other pivotal processes [41]

was repeated for a total of twenty times with each iteration adding one more dataset with samples being derived from a different primary sample site. The non-cancerous samples were all grouped into one class while the cancerous samples from different primary sample sites were grouped into separate classes. The iterations were completed by adding first breast followed by kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts, cervix uteri, soft tissue, retroperitoneal and peritoneum, esophagus, adrenal gland, pancreas, and finally testis. After the twenty iterations were completed, a 2D and 3D PCA was visualized on the last iteration dataset. Finally, using all the available data, all cancerous samples were grouped into one class and all noncancerous samples were grouped into another. The mutual information top-20 feature selection was calculated again on this dataset and the features were used to train two binary classifiers using only the breast cancer data and thyroid cancer data.

III. RESULTS

The following subsections report the results of the experiments outlined in the methods section. The results show the outcomes of all twenty iterations as well as two additional experiments that were conducted on the twentieth iteration dataset.

A. RETRIEVAL OF CLINICALLY RELEVANT MIRNAS

Table 2 shows the best characterized (clinical) cancer-associated miRNAs along with their functionality and biological role identified through literature. The miRNA biomarkers are based on two review studies that identified well characterized miRNAs with extensively documented associations to cancer development [28], [29]. To further confirm the robustness of each miRNA biomarker, we determined the biological role of the miRNA biomarkers identified through the review articles, with the associated study shown in the table. The biological functionality studies affirmed that the well characterized miRNA biomarkers had associations with general, non-tissue specific cancer development processes. Fig. 1 shows the biologically derived miRNA biomarkers identified through literature. These miRNA biomarkers are based on four review studies that comprehensively identified all miRNAs that have been identified as cancer biomarkers in previous research [27], [30], [31], [32]. All biomarkers were identified through biological-based methodologies such as the analysis techniques outlined in subsection G of materials and methods.

let7a	let7b	let7c	let7d	let7e	let7f
let7i	miR- 27a	miR-1	miR-10	miR-100	miR-106a
miR-106b	miR-107	miR-10a	miR-10b	miR-122	miR-1224-3p
miR-1236	miR-1248b	miR-1255b-5p	miR-125a	miR-125b	miR-126
miR-127	miR-127-3p	miR-128	miR-132	miR-133a	miR-133b
miR-135b	miR-141	miR-144-5p	miR-145	miR-146a	miR-146b
miR-148a	miR-150	miR-152	miR-155	miR-15a	miR-15b
miR-16	miR-16-1	miR-17	miR-17-3p	miR-17-5p	miR-17-92
miR-181a-2	miR-181b	miR-183	miR-184	miR-187	miR-18a
miR-19a	miR-191	miR-192	miR-193a	miR-193b	miR-195
miR-196a	miR-197	miR-199a	miR-19a	miR-19b	miR-200a
miR-200b	miR-200c	miR-203	miR-205	miR-206	miR-20a
miR-20b	miR-21	miR-210	miR-212	miR-218	miR-22
miR-221	miR-222	miR-223	miR-224	miR-23b	miR-25
miR-26a	miR-26b-5p	miR-27a	miR-29a	miR-29b	miR-29c
miR-30a	miR-30a-3p	miR-30c	miR-30d	miR-30e	miR-30e-3p
miR-31	miR-32	miR-33	miR-335	miR-338	miR-34
miR-342	miR-34a	miR-34b	miR-34c	miR-373	miR-374-5p
miR-374a	miR-375	miR-378	miR-409	miR-423-5p	miR-425
miR-429	miR-483-5p	miR-486	miR-486-5p	miR-499	miR-500
miR-516a	miR-519d	miR-520b	miR-542-5p	miR-565	miR-574-3p
miR-589	miR-601	miR-618	miR-652	miR-7	miR-760
miR-767-3p	miR-801	miR-891b	miR-9	miR-92	miR-92a
miR-92b	miR-93	miR-95	miR-96	miR-99b	

FIGURE 1. Experimentally identified miRNA Biomarkers [30], [31], [32], and [27].

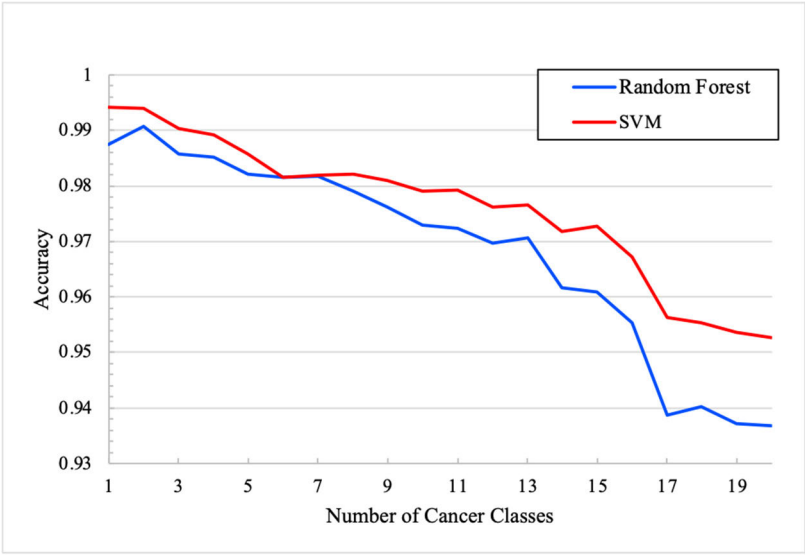

FIGURE 2. Accuracy vs number of cancer classes.

B. CLASSIFICATION MODEL SUCCESS METRICS

Fig. 2-4 show the accuracy, precision, and recall as a function of the number of classes included in the classification model. The figures show that overall, across all metrics, as the number of classes increases, the metrics decrease. While this trend is generally true, there are exceptions at certain iterations. Furthermore, the results show that the models in all iterations were able to classify the data to a high degree

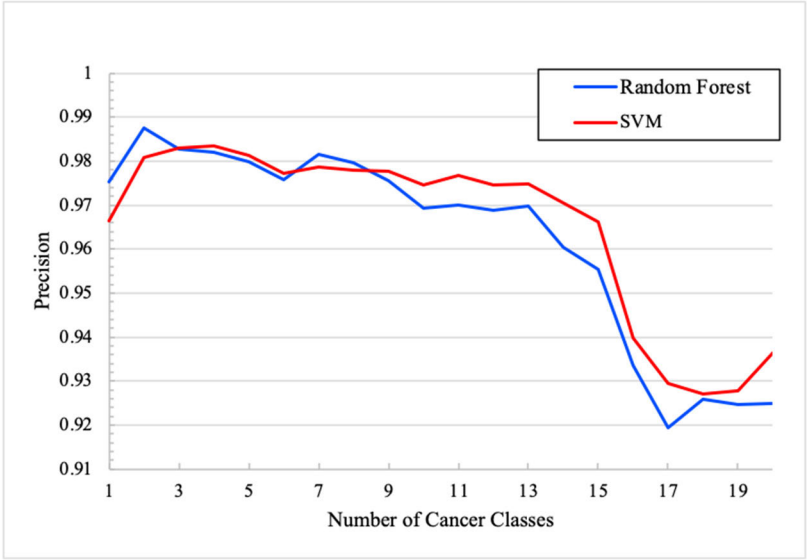

FIGURE 3. Precision vs number of cancer classes.

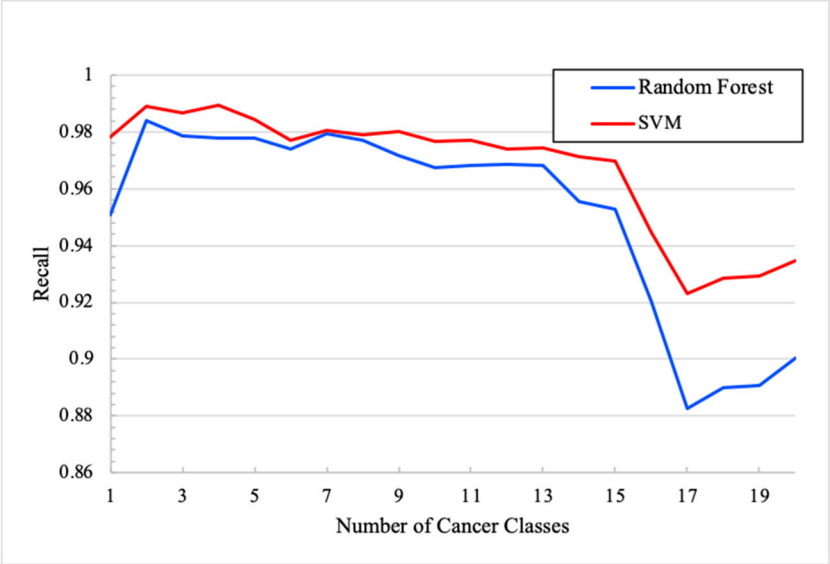

FIGURE 4. Recall vs number of cancer classes.

of success. The random forest model at the final iteration achieved a 0.9367, 0.9249, and 0.9002 accuracy, precision, and recall respectively. The SVM model at the final iteration achieved a 0.9527, 0.9366, and 0.9348 accuracy, precision, and recall respectively. Finally, the SVM performed better than the random forest model on most iterations across all three metrics. From now on, we will focus on using SVM as a classification tool for our investigation unless it is clearly stated otherwise.

C. FEATURE SELECTION VS RELEVANT MIRNAS

Table 3 shows the top-20 features identified for each iteration through feature selection. The features were also compared to the miRNAs identified in the relevant literature to compute the percentage relevance of the features identified as described in subsection G of materials and methods. Fig. 5 shows the relevance as a function of the number of classes included in the classification model. The results show that both biological and clinical relevance increased from the first iteration to the last iteration, starting at 50% biologically relevant and 25% clinically relevant, and ending at 65% and 35% respectively. Relevance (biological and clinical) is defined in subsection G of materials and methods and subsection A of results, with biological relevance utilizing the miRNAs identified in Fig. 1 and clinical relevance utilizing the miRNAs identified in Table 2. Although Fig. 5

TABLE 3. Feature selection across different cancers sets.

Type of Cancers	miRNAs from Top-20 Feature Selection	Biologically Relevant (%)	Clinically Relevant (%)
breast	hsa-let-7c, hsa-mir-100, hsa-mir-10b, hsa-mir-125b-1, hsa-mir-125b-2, hsa-mir-139, hsa-mir-141, hsa-mir-145, hsa-mir-182, hsa-mir-183, hsa-mir-200a, hsa-mir-204, hsa-mir-21, hsa-mir-337, hsa-mir-429, hsa-mir-486-1, hsa-mir-486-2, hsa-mir-592, hsa-mir-96, hsa-mir-99a	50	25
breast, kidney	hsa-mir-10b, hsa-mir-122, hsa-mir-141, hsa-mir-149, hsa-mir-182, hsa-mir-183, hsa-mir-190b, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-199a-2, hsa-mir-199b, hsa-mir-200a, hsa-mir-200c, hsa-mir-203a, hsa-mir-204, hsa-mir-205, hsa-mir-375, hsa-mir-6510, hsa-mir-96	50	15
breast, kidney, corpus uteri	hsa-mir-122, hsa-mir-139, hsa-mir-141, hsa-mir-149, hsa-mir-182, hsa-mir-183, hsa-mir-190b, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-203a, hsa-mir-204, hsa-mir-205, hsa-mir-375, hsa-mir-429, hsa-mir-6510, hsa-mir-96	55	25
breast, kidney, corpus uteri, thyroid gland	hsa-let-7i, hsa-mir-10a, hsa-mir-135a-1, hsa-mir-135a-2, hsa-mir-141, hsa-mir-182, hsa-mir-183, hsa-mir-190b, hsa-mir-196a-1, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-204, hsa-mir-205, hsa-mir-429, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3	55	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung	hsa-let-7i, hsa-mir-10a, hsa-mir-10b, hsa-mir-135a-1, hsa-mir-135a-2, hsa-mir-135b, hsa-mir-141, hsa-mir-182, hsa-mir-183, hsa-mir-190b, hsa-mir-196a-1, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200c, hsa-mir-204, hsa-mir-205, hsa-mir-29a, hsa-mir-375, hsa-mir-429	60	30

TABLE 3. *(Continued.)* **Feature selection across different cancers sets.**

breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland	hsa-let-7i, hsa-mir-10a, hsa-mir-10b, hsa-mir-135a-1, hsa-mir-135a-2, hsa-mir-135b, hsa-mir-141, hsa-mir-181a-1, hsa-mir-181a-2, hsa-mir-181b-1, hsa-mir-181b-2, hsa-mir-183, hsa-mir-192, hsa-mir-196a-1, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200c, hsa-mir-204, hsa-mir-205, hsa-mir-375	55	15
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain	hsa-mir-10a, hsa-mir-10b, hsa-mir-135a-1, hsa-mir-135a-2, hsa-mir-135b, hsa-mir-141, hsa-mir-181a-2, hsa-mir-196a-1, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-199b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-205, hsa-mir-375, hsa-mir-429, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3	65	30
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary	hsa-mir-10a, hsa-mir-126, hsa-mir-141, hsa-mir-153-2, hsa-mir-181a-2, hsa-mir-196a-1, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-199a-2, hsa-mir-199b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-21, hsa-mir-375, hsa-mir-429, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-mir-92b	65	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach	hsa-mir-10a, hsa-mir-141, hsa-mir-181a-2, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-199b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3	65	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon	hsa-mir-10a, hsa-mir-141, hsa-mir-181a-2, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-1, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3	65	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin	hsa-mir-10a, hsa-mir-135a-1, hsa-mir-141, hsa-mir-181a-2, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3	65	35

TABLE 3. *(Continued.)* **Feature selection across different cancers sets.**

breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder	hsa-mir-10a, hsa-mir-141, hsa-mir-153-2, hsa-mir-181a-2, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3	65	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts	hsa-mir-10a, hsa-mir-141, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-885, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-mir-92b	65	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts, cervix uteri	hsa-mir-10a, hsa-mir-141, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-885, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-mir-92b	65	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts, cervix uteri, soft tissue	hsa-mir-10a, hsa-mir-141, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-203a, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-mir-92b	65	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts, cervix uteri, soft tissue, retroperitoneal and peritoneum	hsa-mir-141, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-203a, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-885, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-mir-92b	60	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts, cervix uteri, soft tissue, retroperitoneal and peritoneum, esophagus	hsa-mir-141, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-203a, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-885, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-mir-92b	60	35

TABLE 3. *(Continued.)* **Feature selection across different cancers sets.**

breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts, cervix uteri, soft tissue, retroperitoneal and peritoneum, esophagus, adrenal gland	hsa-mir-10a, hsa-mir-141, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196a-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-203a, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-mir-92b	65	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts, cervix uteri, soft tissue, retroperitoneal and peritoneum, esophagus, adrenal gland, pancreas	hsa-mir-10a, hsa-mir-141, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-203a, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-885, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-mir-92b	65	35
breast, kidney, corpus uteri, thyroid gland, bronchus and lung, prostate gland, brain, ovary, stomach, colon, skin, bladder, liver and intrahepatic bile ducts, cervix uteri, soft tissue, retroperitoneal and peritoneum, esophagus, adrenal gland, pancreas, testis	hsa-mir-10a, hsa-mir-141, hsa-mir-192, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-196b, hsa-mir-200a, hsa-mir-200b, hsa-mir-200c, hsa-mir-203a, hsa-mir-205, hsa-mir-21, hsa-mir-215, hsa-mir-375, hsa-mir-429, hsa-mir-885, hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-mir-92b	65	35

does show a slight upward trend in the percentage relevance, the increase eventually stabilizes as the number of classes increases for both biological and clinical percentage relevance.

D. PCA ANALYSIS AND VISUALIZATION

Fig. 6 shows the results of a 2D PCA on the dataset containing 20 cancer classes (i.e., full dataset) and Fig. 7a-7c show the results of a 3D PCA on the dataset from different viewing angles. The PCAs show the relative space occupied by each class after reducing the dimensionality of the feature space to 2 and 3, respectively. The PCA graphs show that the space occupied by the primary sample sites varied. Each sample site had a specific pattern of miRNA expression that led to distinct areas within the PCA. Furthermore, all the non-cancerous samples, regardless of the primary sample site, occupied a distinct area near the origin of the PCA visualizations. This is more clearly seen on the 3D PCA visualizations where all the non-cancerous samples create a spine-like area near the origin. The cancerous samples branch out from this area in distinct ways.

E. TOP 20 FEATURE SELECTION ON THE FULL DATASET FOR BINARY CLASSIFICATION

Table 4 and Fig. 8 show the results of the top 20 selected features executed on the full dataset as a binary classification problem, in which all cancer samples from different cancer classes were grouped into one and all non-cancer samples are grouped into the other. The 20 miRNAs identified through the feature selection are shown in Fig. 8. The performance metrics of the binary classification test using only the 20 miRNAs identified are shown in Table 4 for two cancer types. The results show that the binary breast cancer classification model achieved an accuracy, precision, and recall of 0.9892, 0.9645, and 0.9585 respectively with the SVM model. Additionally, the binary thyroid cancer classification model achieved an accuracy, precision, and recall of 0.9669, 0.9152, and 0.9164 respectively for the SVM model.

IV. DISCUSSIONS

This study focused on exploring characteristics of creating a multi-cancer diagnostic model using miRNA expression profiles. Previous machine learning approaches to miRNA-based

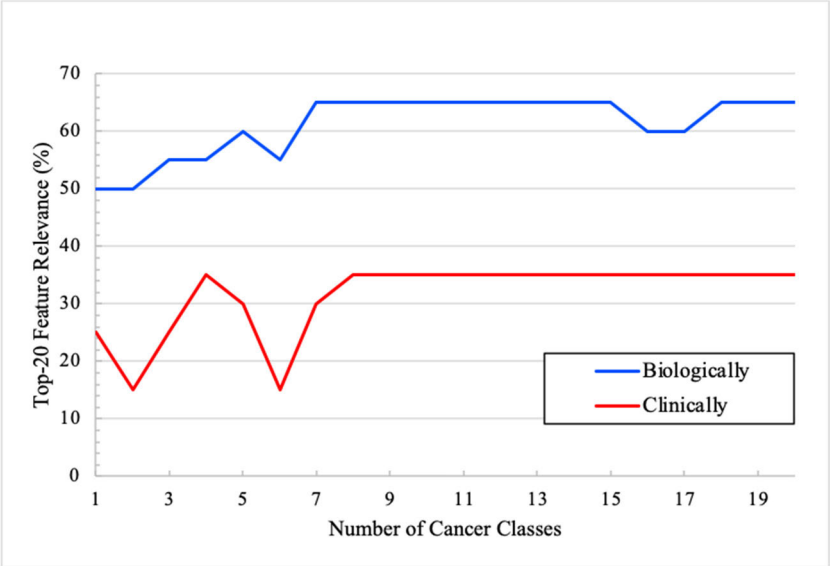

FIGURE 5. Relevance vs number of cancer classes.

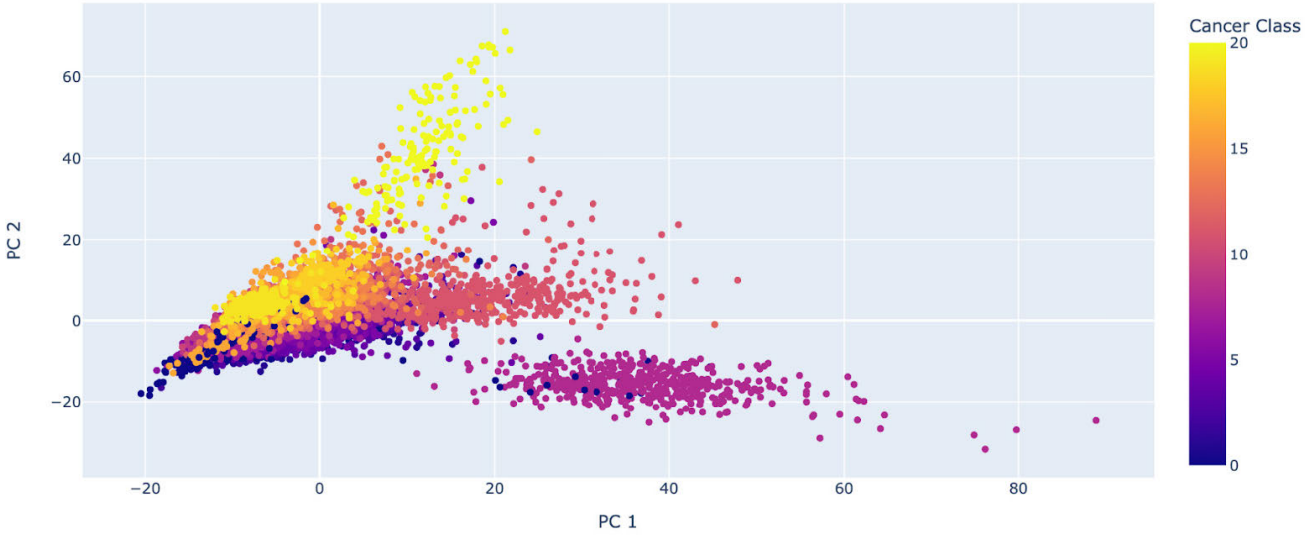

FIGURE 6. 2-Dimensional PCA of the full dataset.

TABLE 4. Binary classification results based on Top-20 features of the full dataset.

Type of Cancer	SVM Accuracy	SVM Precision	SVM Recall
Breast	0.9892	0.9645	0.9585
Thyroid	0.9669	0.9152	0.9164

cancer classification have identified numerous miRNAs through feature extraction techniques as potential targets for biological research [20]. These miRNAs ranked highly during feature extraction processes, identifying them as candidates for potential biomarkers. However, many of them do not have any noteworthy biological role or significance, making them

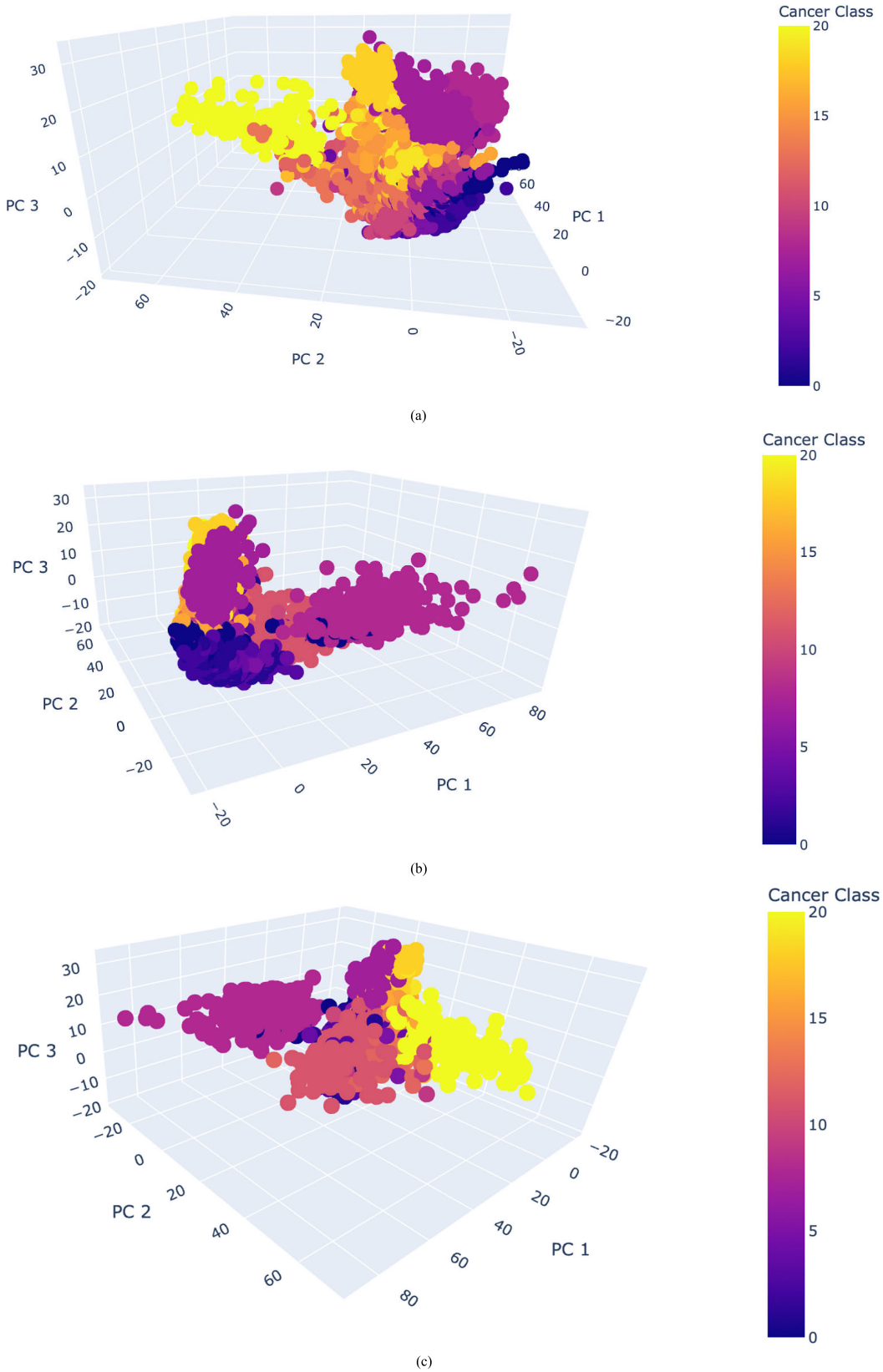

FIGURE 7. a 3-Dimensional PCA of the full dataset. b Another view of the 3-Dimensional PCA of the Full Dataset. c Yet Another view of the 3-Dimensional PCA of the Full Dataset.

hsa-mir-106b	hsa-mir-10b	hsa-mir-1258	hsa-mir-1301	hsa-mir-130b	hsa-mir-139
hsa-mir-141	hsa-mir-145	hsa-mir-182	hsa-mir-183	hsa-mir-195	hsa-mir-21
hsa-mir-210	hsa-mir-301b	hsa-mir-474	hsa-mir-769	hsa-mir-7706	hsa-mir-93
hsa-mir-937	hsa-mir-96				

FIGURE 8. Top-20 features of the full dataset.

improbable for use in any approved clinical applications. miRNAs have significant promise for future diagnostic tests because they can be detected directly from biological fluids, such as blood, urine, saliva, and pleural fluid, as well as the availability of high-quality measurement techniques for miRNAs [27]. This makes understanding and characterizing the biological basis behind potential miRNA classification tools crucial for integration into clinical environments.

In this research, we attempt to iteratively classify cancer types and link leading miRNAs identified with a data science approach to those with clinical and biological means. Firstly, the results show a clear decrease in accuracy, precision, and recall as more classes are added to the machine learning model. This is an expected trend when increasing the number of classes in a multiclass classifier, as more cancer classes will introduce more complex miRNA expression patterns to model. Thus, the graphs illustrate a steady decline in performance as the number of classes are increased. Notably, the SVM did perform better than the random forest overall, ending at an accuracy of around 95% while the random forest had an accuracy of around 94% by the 20-cancer-class model. However, the differences between the SVM and random forest are marginal for our purposes, and importantly both models follow the same trend in performance as the number of classes increase. Regarding clinically relevant miRNAs, the percentage of relevance increased from the first iteration to the last iteration for both clinical and biological relevance among both the SVM and random forest. This means that the decrease in model success associated with the increase in classes does not parallel the trend in relevance. Thus, no correlation can be established between a decrease in success and a decrease in relevance. However, the increase in relevance may be due to the increased generalizability of the 20-miRNA feature extraction signature as the model adapts to classify more cancer classes.

The list of biologically relevant miRNAs is a comprehensive record from robust studies that identify common miRNA biomarkers in multiple studies. The studies attempted to comprehensively identify potential miRNA biomarkers experimentally recognized through established biological analysis procedures; however, many are not clinically relevant, and associations may not be well established to cancer. On the other hand, the list of clinically relevant miRNAs represents only well characterized cancer associated miRNAs, many of which are involved in clinical applications and therapies. These miRNAs have well studied roles and relationships between the miRNA and function/role in cancer across multiple studies. Additionally, the miRNAs in the clinically relevant table serve a diverse range of cancerous tissues and represent general biological roles important for cancer formation such as cell metabolism, proliferation, differentiation, and apoptosis. Thus, the trend of greater percentage relevance as more cancer classes are added may be attributed to an increase in model focus on such broad-scope clinically relevant miRNAs. Initially, when the model only needed to classify a single tissue, it would focus on specific patterns relevant to only that cancer type. However, a larger set of cancers rendered such single tissue specific miRNAs less important, focusing more on miRNAs variably expressed by all cancer classes, such as those identified in the clinical and biological relevance table. We found in this study that the maximum relevance occurred at 65% for biological relevance and 35% for clinical relevance. The results show that models use significantly more than just biologically identified miRNAs, illustrating that more research is needed on biomarker miRNAs identified through data science techniques without established biological research.

To investigate if the feature extraction signature was becoming more generalizable as the number of classes increased, we used the 20-miRNA feature extraction signature from the full-dataset binary model to create a binary classifier for breast and thyroid cancer. The models performed with a rough accuracy of around 96% and 99% across both models for thyroid and breast cancer respectively, showing the feasibility of using this multi-cancer signature as a general cancer signature for a single cancer type. The 20-miRNA feature extraction signature was established using 20 different classes of cancer combined as a single cancer and a single non-cancer class. This allowed the model to focus on miRNAs that had similar expressions among all cancerous tissues and among all non-cancerous tissues but differed when comparing any one cancerous tissue to a non-cancerous tissue. Thus, the success of using this general cancer miRNA signature to classify a specific tissue type as cancerous or non-cancerous shows that there are similarities in the change in miRNA expression that occurs when any tissue becomes cancerous.

Finally, our PCA graph of the 20 cancer classes showed the distinct separability of each type from one another. Notably, all non-cancerous data was similarly clustered closer to the

origin of the PCA graphs, illustrating that all non-cancerous tissues regardless of origin may have similar expressions of important miRNA features. The specific areas occupied by each primary sample site show that each tissue-specific cancer type changes its miRNA expression profile in distinct ways. This may indicate that cancers from different tissues impact cells in different ways, but cancer among a specific tissue type follows a pattern. The differentiated space shown in the PCA visualizations also highlight the differentiability of the classes that allowed for the high degree of classification success shown across all iterations.

Although our study demonstrated trends in miRNA relevance and accuracy through an iterative approach of increasing cancer classes, our data was based on the TCGA dataset from the GDC data portal. The data was notably unbalanced, with some classes having no non-cancerous tissue expression data points. Additionally, many tissue classes had relatively few data points as the iterations increased compared to the initial classes, as shown in Table 1. Additionally, when conducting the literature review to identify miRNA biomarkers for our relevance comparison, the methods used to analyze and identify biomarkers differed between studies. Studies have shown that the choice in analysis technique can influence expression measurements to the point where a lack of correspondence between platforms have been caused when using the same sample source [27]. Thus, the identified biomarkers may have a degree of variability, causing us to potentially omit identified biomarkers due to different quantification standards used in different studies.

V. CONCLUSION

This study explored the relationship between the relevant miRNAs identified through feature selection and the performance metrics of the classification models across twenty iterations. Each iteration added another primary sample site to the multi-class models, increasing the number of cancer types involved. The results showed that despite a decrease in performance metrics across the iterations, the twenty cancer types can be classified to a high degree of success. Furthermore, the relevance increased from the first iteration to the last iteration for both biological and clinical relevance. This shows that as more cancer classes are involved, the model generalized towards cancer as a whole. The PCA also showed that each class occupied a distinct spatial region. Further research is needed to explore the significance of the miRNAs that were not considered biologically and clinically relevant but were still identified in the feature selection. The relationship between model generalizability, expression differences between cancer types, and model performance can be further investigated using different datasets without the limitations discussed.

AUTHOR CONTRIBUTIONS

Matthew Acs and Richard Acs contributed the study design and implementation, study selection, manuscript drafting, manuscript revisions, and a literature review of miRNA biomarkers. Matthew Acs and Richard Acs provided equal contributions. Charles Briandi and Eyan Eubanks contributed to the data analysis and codebase. Oneeb Rehman and Hanqi Zhuang contributed to the analysis of results and manuscript revisions.

CODE AVAILABILITY

The source code for the experiments is available as Jupyter Notebooks on GitHub at https://github.com/Mattliketocode/miRNA-Cancer-Classification.

ACKNOWLEDGMENT

The authors would like to thank the National Cancer Institute for providing the genomic data commons data portal that contained the TCGA data that we used and also would like to thank Google Colab for providing cloud hosted Jupyter Notebooks for the codebase.

(Matthew Acs and Richard Acs are co-first authors.)

REFERENCES

[1] L. Tomasello, R. Distefano, G. Nigita, and C. M. Croce, "The MicroRNA family gets wider: The IsomiRs classification and role," *Frontiers Cell Develop. Biol.*, vol. 9, Jun. 2021, Art. no. 668648, doi: 10.3389/fcell.2021.668648.

[2] T.-Y. Ha, "MicroRNAs in human diseases: From cancer to cardiovascular disease," *Immune Netw.*, vol. 11, no. 3, p. 135, 2011, doi: 10.4110/in.2011.11.3.135.

[3] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, "MicroRNAs and complex diseases: From experimental results to computational models," *Briefings Bioinf.*, vol. 20, no. 2, pp. 515–539, Oct. 2017, doi: 10.1093/bib/bbx130.

[4] N. J. Meropol and K. A. Schulman, "Cost of cancer care: Issues and implications," *J. Clin. Oncol.*, vol. 25, no. 2, pp. 180–186, Jan. 2007, doi: 10.1200/jco.2006.09.6081.

[5] A. Esquela-Kerscher and F. J. Slack, "Oncomirs—microRNAs with a role in cancer," *Nature Rev. Cancer*, vol. 6, no. 4, pp. 259–269, Apr. 2006, doi: 10.1038/nrc1840.

[6] Y. Piao, M. Piao, and K. H. Ryu, "Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles," *Comput. Biol. Med.*, vol. 80, pp. 39–44, Jan. 2017, doi: 10.1016/j.compbiomed.2016.11.008.

[7] S. W. Fu, L. Chen, and Y.-G. Man, "miRNA biomarkers in breast cancer detection and management," *J. Cancer*, vol. 2, pp. 116–122, Jan. 2011, doi: 10.7150/jca.2.116.

[8] X. Zhu, M. Kudo, X. Huang, H. Sui, H. Tian, C. M. Croce, and R. Cui, "Frontiers of MicroRNA signature in non-small cell lung cancer," *Frontiers Cell Develop. Biol.*, vol. 9, Apr. 2021, Art. no. 643942, doi: 10.3389/fcell.2021.643942.

[9] S. Ghafouri-Fard, Z. Shirvani-Farsani, W. Branicki, and M. Taheri, "MicroRNA signature in renal cell carcinoma," *Frontiers Oncol.*, vol. 10, Nov. 2020, Art. no. 596359, doi: 10.3389/fonc.2020.596359.

[10] X. W. Wang, N. H. H. Heegaard, and H. Ørum, "MicroRNAs in liver disease," *Gastroenterology*, vol. 142, no. 7, pp. 1431–1443, Jun. 2012, doi: 10.1053/j.gastro.2012.04.007.

[11] H. Ogata-Kawata, M. Izumiya, D. Kurioka, Y. Honma, Y. Yamada, K. Furuta, T. Gunji, H. Ohta, H. Okamoto, H. Sonoda, M. Watanabe, H. Nakagama, J. Yokota, T. Kohno, and N. Tsuchiya, "Circulating exosomal microRNAs as biomarkers of colon cancer," *PLoS ONE*, vol. 9, no. 4, Apr. 2014, Art. no. e92921, doi: 10.1371/journal.pone.0092921.

[12] (2019). *GDC*. Cancer.gov. [Online]. Available: https://portal.gdc.cancer.gov/

[13] GEO. (2019). *Home-GEO-NCBI*. Nih.gov. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/

[14] K. Kalecky, R. Modisette, S. Pena, Y.-R. Cho, and J. Taube, "Integrative analysis of breast cancer profiles in TCGA by TNBC subgrouping reveals novel microRNA-specific clusters, including miR-17-92a, distinguishing basal-like 1 and basal-like 2 TNBC subtypes," *BMC Cancer*, vol. 20, no. 1, pp. 1–13, Feb. 2020, doi: 10.1186/s12885-020-6600-6.

[15] Z. Yang, H. Yin, L. Shi, and X. Qian, "A novel microRNA signature for pathological grading in lung adenocarcinoma based on TCGA and GEO data," *Int. J. Mol. Med.*, vol. 45, no. 5, pp. 1397–1408, Mar. 2020, doi: 10.3892/ijmm.2020.4526.

[16] C.-Y. Li, W.-W. Zhang, J.-L. Xiang, X.-H. Wang, J. Li, and J.-L. Wang, "Identification of microRNAs as novel biomarkers for esophageal squamous cell carcinoma," *Chin. Med. J.*, vol. 132, no. 18, pp. 2213–2222, Sep. 2019, doi: 10.1097/cm9.0000000000000427.

[17] H. Jacob, L. Stanisavljevic, K. E. Storli, K. E. Hestetun, O. Dahl, and M. P. Myklebust, "A four-microRNA classifier as a novel prognostic marker for tumor recurrence in stage II colon cancer," *Sci. Rep.*, vol. 8, no. 1, p. 6157, Apr. 2018, doi: 10.1038/s41598-018-24519-4.

[18] Y. M. Youssef, N. M. A. White, J. Grigull, A. Krizova, C. Samy, S. Mejia-Guerrero, A. Evans, and G. M. Yousef, "Accurate molecular classification of kidney cancer subtypes using MicroRNA signature," *Eur. Urol.*, vol. 59, no. 5, pp. 721–730, May 2011, doi: 10.1016/j.eururo.2011.01.004.

[19] A. G. Telonis, R. Magee, P. Loher, I. Chervoneva, E. Londin, and I. Rigoutsos, "Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types," *Nucleic Acids Res.*, vol. 45, no. 6, pp. 2973–2985, Feb. 2017, doi: 10.1093/nar/gkx082.

[20] A. Lopez-Rincon, M. Martinez-Archundia, G. U. Martinez-Ruiz, A. Schoenhuth, and A. Tonda, "Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection," *BMC Bioinf.*, vol. 20, no. 1, pp. 1–17, Sep. 2019, doi: 10.1186/s12859-019-3050-8.

[21] N. Cheerla and O. Gevaert, "MicroRNA based pan-cancer diagnosis and treatment recommendation," *BMC Bioinf.*, vol. 18, no. 1, pp. 1–11, Jan. 2017, doi: 10.1186/s12859-016-1421-y.

[22] J. Laplante and M. A. Akhloufi, "Predicting cancer types from miRNA stem-loops using deep learning," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 5312–5315, doi: 10.1109/embc44109.2020.9176345.

[23] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138, doi: 10.1103/physreve.69.066138.

[24] I. Tougui, A. Jilbab, and J. E. Mhamdi, "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications," *Healthcare Informat. Res.*, vol. 27, no. 3, pp. 189–199, Jul. 2021, doi: 10.4258/hir.2021.27.3.189.

[25] M. Consortium, "The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nature Biotechnol.*, vol. 28, no. 8, pp. 827–838, Jul. 2010, doi: 10.1038/nbt.1665.

[26] Y.-H. Taguchi and Y. Murakami, "Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers," *PLoS ONE*, vol. 8, no. 6, Jun. 2013, Art. no. e66714, doi: 10.1371/journal.pone.0066714.

[27] E. Larrea, C. Sole, L. Manterola, I. Goicoechea, M. Armesto, M. Arestin, M. Caffarel, A. Araujo, M. Araiz, M. Fernandez-Mercado, and C. Lawrie, "New concepts in cancer biomarkers: Circulating miRNAs in liquid biopsies," *Int. J. Mol. Sci.*, vol. 17, no. 5, p. 627, Apr. 2016, doi: 10.3390/ijms17050627.

[28] G. D. Leva and C. M. Croce, "miRNA profiling of cancer," *Current Opinion Genet. Develop.*, vol. 23, no. 1, pp. 3–11, Feb. 2013, doi: 10.1016/j.gde.2013.01.004.

[29] J. Hayes, P. P. Peruzzi, and S. Lawler, "MicroRNAs in cancer: Biomarkers, functions and therapy," *Trends Mol. Med.*, vol. 20, no. 8, pp. 460–469, Aug. 2014, doi: 10.1016/j.molmed.2014.06.005.

[30] J. Wang, K.-Y. Zhang, S.-M. Liu, and S. Sen, "Tumor-associated circulating MicroRNAs as biomarkers of cancer," *Molecules*, vol. 19, no. 2, pp. 1912–1938, Feb. 2014, doi: 10.3390/molecules19021912.

[31] G. Cheng, "Circulating miRNAs: Roles in cancer diagnosis, prognosis and therapy," *Adv. Drug Del. Rev.*, vol. 81, pp. 75–93, Jan. 2015, doi: 10.1016/j.addr.2014.09.001.

[32] K. Sundarbose, R. Kartha, and S. Subramanian, "MicroRNAs as biomarkers in cancer," *Diagnostics*, vol. 3, no. 1, pp. 84–104, Jan. 2013, doi: 10.3390/diagnostics3010084.

[33] B. Boyerinas, S.-M. Park, A. Hau, A. E. Murmann, and M. E. Peter, "The role of let-7 in cell differentiation and cancer," *Endocrine-Related Cancer*, vol. 17, no. 1, pp. F19–F36, Mar. 2010, doi: 10.1677/ERC-09-0184.

[34] M. J. Ramaiah, "Functions and epigenetic aspects of miR-15/16: Possible future cancer therapeutics," *Gene Rep.*, vol. 12, pp. 149–164, Sep. 2018, doi: 10.1016/j.genrep.2018.06.012.

[35] A. J. Kriegel, Y. Liu, Y. Fang, X. Ding, and M. Liang, "The miR-29 family: Genomics, cell biology, and relevance to renal and cardiovascular injury," *Physiolog. Genomics*, vol. 44, no. 4, pp. 237–244, Feb. 2012, doi: 10.1152/physiolgenomics.00141.2011.

[36] L. Zhang, Y. Liao, and L. Tang, "MicroRNA-34 family: A potential tumor suppressor and therapeutic candidate in cancer," *J. Experim. Clin. Cancer Res.*, vol. 38, no. 1, pp. 1–13, Feb. 2019, doi: 10.1186/s13046-019-1059-5.

[37] C. Li, Y. Li, Y. Lu, Z. Niu, H. Zhao, Y. Peng, and M. Li, "MiR-26 family and its target genes in tumorigenesis and development," *Crit. Rev. Oncol./Hematol.*, vol. 157, Jan. 2021, Art. no. 103124, doi: 10.1016/j.critrevonc.2020.103124.

[38] P. M. Neilsen, J. E. Noll, S. Mattiske, C. P. Bracken, P. A. Gregory, R. B. Schulz, S. P. Lim, R. Kumar, R. J. Suetani, G. J. Goodall, and D. F. Callen, "Mutant p53 drives invasion in breast tumors through up-regulation of miR-155," *Oncogene*, vol. 32, no. 24, pp. 2992–3000, Jul. 2012, doi: 10.1038/onc.2012.305.

[39] R. Kumarswamy, I. Volkmann, and T. Thum, "Regulation and function of miRNA-21 in health and disease," *RNA Biol.*, vol. 8, no. 5, pp. 706–713, Sep. 2011, doi: 10.4161/rna.8.5.16154.

[40] J. Song, Y. Ouyang, J. Che, X. Li, Y. Zhao, K. Yang, X. Zhao, Y. Chen, C. Fan, and W. Yuan, "Potential value of miR-221/222 as diagnostic, prognostic, and therapeutic biomarkers for diseases," *Frontiers Immunol.*, vol. 8, p. 56, Feb. 2017, doi: 10.3389/fimmu.2017.00056.

[41] E. Mogilyansky and I. Rigoutsos, "The miR-17/92 cluster: A comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease," *Cell Death Differentiation*, vol. 20, no. 12, pp. 1603–1614, Nov. 2013, doi: 10.1038/cdd.2013.125.

MATTHEW ACS is currently pursuing the Bachelor of Science degree in computer science with Florida Atlantic University (FAU). He plans to complete the Master of Science degree in artificial intelligence and the Ph.D. degree in computer science with FAU following his graduate. His research interests include applied machine learning, generative AI, XAI, and image classification and segmentation in biology and medicine.

RICHARD ACS is currently pursuing the Bachelor of Science degree in computer science with Florida Atlantic University (FAU). He plans to complete the master's degree in artificial intelligence and the Ph.D. degree in computer science with FAU after his undergraduate degree. His research interests include machine learning in medical domains and video processing using deep learning.

CHARLES BRIANDI is from Venice, FL, USA. He is currently pursuing the bachelor's degree in computer science (artificial intelligence) with Florida Atlantic University (FAU). He attended the Venice High School. Since high school, he has been interested in software developing, where he began learning basic programming in C from online resources. After graduating, in 2019, he attended FAU. His research interests include artificial intelligence and backend development.

EYAN EUBANKS is currently pursuing the Bachelor of Science degree in computer science with Florida Atlantic University. His current research interests include machine learning and cybersecurity. In the future, he plans to further his studies with the Master of Science degree in machine learning. His passion is to learn about machine learning and how it can be applied to cybersecurity.

ONEEB REHMAN is currently pursuing the Ph.D. degree in electrical engineering with Florida Atlantic University (FAU). He is currently a Graduate Teaching Assistant with FAU. His research interests include bioinformatics, parallel computing, and cancer detection and classification.

HANQI ZHUANG (Senior Member, IEEE) is currently a Professor in electrical engineering with Florida Atlantic University (FAU), Boca Raton. He has received research grants from various federal agencies and local industries. He has published approximately 50 papers in refereed international journals and has given numerous presentations in conferences and institutions. He has been greatly involved in developing and delivering, sometimes through remote means, cross-disciplinary courses, and laboratories with support mainly from the National Science Foundation. These newly developed courses have received very positive feedback from both undergraduate and graduate students. His research interests include robotics, computer vision, parallel computing, and biometrics. He is on the editorial board of the *International Journal of Computer Applications*. He is currently an Associate Editor of IEEE TRANSACTIONS ON ROBOTICS. He received the FAU undergraduate Teaching of Excellence Award, in 2003, and the Dean's Award, in 2005.