Deep Learning Based Identification of Tissue of Origin for Carcinomas of Unknown Primary utilizing micro-RNA expression

Ananya Raghu, Anisha Raghu, Jillian Wise

Submitted to: JMIR Bioinformatics and Biotechnology
on: January 18, 2024

Table of Contents

Deep Learning Based Identification of Tissue of Origin for Carcinomas of Unknown Primary utilizing micro-RNA expression

Ananya Raghu[1*]; Anisha Raghu[1*]; Jillian Wise[2]

[1]Quarry Lane School San Ramon US
[2]Salve Regina University Newport US
[*]these authors contributed equally

Corresponding Author:
Jillian Wise
Salve Regina University
100 Ochre Point Ave
Newport
US

Abstract

Background: Carcinoma of Unknown Primary (CUP) is a subset of metastatic cancers in which the primary tissue source of the cancer cells, or origin, remains unidentified. CUP is the eighth most common malignancy worldwide, and accounts for up to five percent of all malignancies. Representing an exceptionally aggressive category of metastatic cancers, the median survival of CUP is approximately three to six months. The tissue in which a cancer arises plays a key role in our understanding of sensitivities to various forms of cell death in cancer cells. Thus, the lack of knowledge on tissue of origin makes it difficult to devise tailored and effective treatments for patients with CUP. Developing quick and clinically implementable methods to identify the tissue of origin of the primary site is crucial in treating CUP patients. Non-coding RNAs, may hold potential for origin identification and provide a robust route to clinical implementation due to their resistance against chemical degradation.

Objective: In this work, we investigate the potential of microRNAs, a subset of non-coding RNAs, as highly accurate biomarkers for detecting the tissue of origin through data driven, machine learning, approaches for metastatic cancers.

Methods: We use microRNA expression data from the Cancer Genome Atlas (TCGA) dataset and assess various machine learning approaches, from simple classifiers to deep learning approaches. As a validation of our classifiers, we evaluate the accuracy on a separate set of 194 primary tumor samples from the Sequence Read Archive (SRA). We use permutation feature importance to determine potential miRNA biomarkers and assess with PCA and t-SNE visualizations.

Results: Our results show that it is possible to design robust classifiers to detect the tissue of origin for metastatic samples on the TCGA dataset with an accuracy of up to 96%, which may be utilized in situations of CUP. Our findings demonstrate that deep learning techniques enhance prediction accuracy. We progressed from an initial accuracy prediction of 62.5% with decision trees to 93.2% with logistic regression, finally achieving 96.1% accuracy using deep learning on metastatic samples. On the SRA validation set, a lower accuracy of 41.2% was achieved by decision tree, while deep learning achieved a higher accuracy of 81.2%. Notably, our feature importance analysis showed the top three most important biomarkers for predicting tissue of origin to be mir-10b, mir-205, and mir-196b, which aligns with previous work.

Conclusions: Our findings highlight the potential of using machine learning techniques to devise accurate tests for detecting tissue of origin for CUP. Since microRNAs are carried throughout the body via extracellular vesicles secreted from cells, they may serve as key biomarkers for liquid biopsy due to their presence in blood plasma. Our work serves as a foundation towards developing blood-based cancer detection tests based on microRNA presence.

(JMIR Preprints 18/01/2024:56538)
DOI: https://doi.org/10.2196/preprints.56538

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Deep Learning Based Identification of Tissue of Origin for Carcinomas of Unknown Primary utilizing micro-RNA expression

Ananya Raghu* [1], Anisha Raghu* [2], Jillian F. Wise, PhD[3,4,5]

[1,2] Quarry Lane School, San Ramon, CA

[3] Broad Institute of MIT and Harvard

[4]Tufts University, Pre-College Programs, Medford, MA

[5]Salve Regina University, Newport, RI

Student Authors

Ananya Raghu, Anisha Raghu

**equal contribution authors*

ABSTRACT

Carcinoma of Unknown Primary (CUP) is a subset of metastatic cancers in which the primary tissue source, or origin, remains unidentified. CUP accounts for two to five percent of all malignancies [1]. Representing an exceptionally aggressive category of metastatic cancers, the median survival of those diagnosed with CUP is approximately six to eight months [2]. The tissue in which a cancer arises plays a key role in our understanding of altered gene expression, altered cellular pathways, and sensitivities to various forms of cell death in cancer cells [3]. Thus, the lack of knowledge on tissue of origin (TOO) makes it difficult to devise tailored treatments for patients with CUP. Developing clinically implementable methods to identify the TOO of the primary site is crucial in treating CUP patients [4]. In particular, the expression profiles of non-coding RNAs can provide insight into the TOO for CUP. Non-coding RNAs provide a robust route to clinical implementation due to their resistance against chemical degradation [5].

In this work, we investigate the potential of microRNAs as highly accurate features for detecting the TOO for metastatic cancers. We further hypothesize that data driven approaches can identify specific microRNA features that can determine the TOO. We used microRNA expression data from the Cancer Genome Atlas (TCGA) dataset [6] and assessed various machine learning approaches. Our results show that it is possible to design robust classifiers to detect the TOO for metastatic samples on the TCGA dataset with an accuracy of up to 97%, which may be utilized in situations of CUP. As a validation of our classifiers, we evaluated the accuracy on a separate set of 194 primary tumor samples from the Sequence Read Archive (SRA) [7]. Our findings demonstrate that deep learning techniques enhance prediction accuracy. We progressed from an initial accuracy prediction of 62.5% with decision trees to 94.2% with random forest, finally achieving 97% accuracy using deep learning on metastatic samples. On the SRA test set, a lower accuracy of 41.2% was achieved by decision tree, while deep learning achieved a higher accuracy of 80.4%. Notably, our feature importance analysis showed the top three important miRNA targets for predicting TOO to be mir-10b, mir-205, and mir-196b, which aligns with previous work [8, 9, 10]. Our findings highlight the potential of using machine learning techniques to devise tests for detecting TOO for CUP. Since microRNAs are carried throughout the body via vesicles secreted from cells, future work could investigate their use as key biomarkers for liquid biopsy due to their presence in blood plasma [11]. Our work serves as a foundation towards developing blood-based cancer detection tests based on microRNA presence.

INTRODUCTION

Carcinoma of unknown origin (CUP) occurs when a patient presents at diagnosis with malignant disease across the body, yet the cancer cells tissue of origin (TOO) remains unidentifiable. Thus, CUP is a unique subset of metastasized cancer representing an advanced stage in which the cancer has gained the ability to thrive in new tissue sites and has spread from the primary tumor site. In the United States, an estimated 31490 people were diagnosed with cases of cancer of unknown TOO in 2008. This accounts for nearly 3-5% of all cancer cases and given the lack of knowledge on tissue response to current therapeutics the median survival of patients remains only 3-9 months [12]. In many cases of CUP the primary site is never identified, preventing the use of treatment that can be effective for the true TOO [13]. It has been demonstrated that pinpointing the primary site can significantly increase survival rates by enabling for precise and targeted treatment [14].

Unfortunately, primary tumor identification poses various challenges. Techniques such as serum tumor markers and imaging tests are used to identify the TOO, though only 30% of these tests are successful. Moreover, some positive findings can be misleading [15] and CUP diagnostic workups are often time-consuming, expensive, and unsuccessful [16]. These difficulties have spurred interest in utilizing genetic expression data, such as microRNA, to identify the tissue of origin.

MicroRNAs belong to a class of non-coding regulatory RNAs, small single stranded RNA molecules that are between 19-25 nucleotides long, are involved in the regulation of gene expression of mRNAs. MicroRNAs hold promise as informative biomarkers for cancer due to their significant involvement in cellular processes such as cell division, apoptosis, proliferation and oncogenesis [17]. Beyond their intracellular role in gene regulation, microRNAs may be carried throughout the body via extracellular vesicles secreted from cells and have been identified in the blood. Additionally, miRNA, unlike mRNA, are characterized by resistance to extreme temperatures and pH. This makes microRNAs far more stable biomarkers [18].
Previous work [19] demonstrates that microRNA expression is more informative in classifying tumor samples by their origin in comparison to mRNA. Specifically, microRNAs are better at classifying poorly differentiated tumors [20]. Moreover, microRNAs have shown great potential for identifying tissue of origin for cancers of unknown primary origin [21]. microRNAs have been investigated as prognostic and diagnostic biomarkers extensively in the research community, and have even been found to be deregulated in numerous cancers [22].

With the wide availability of large datasets containing gene expression data, computational techniques such as machine learning have emerged as promising tools for improving TOO detection. Machine learning implementations have increased accuracy in predicting cancer, and have the potential to improve the diagnosis, prognosis, and therapy selection for cancer patients [23]. Three traditional machine learning models are decision trees, random forests, and logistic regression. Decision trees [24] attempt to partition the training set into subsets that contain samples of only one class, thereby predicting the class of interest. Random forests are ensemble classifiers, combining multiple trees for higher accuracy [25]. In contrast, logistic regression is a predictive algorithm to find a model that can predict categorical output [26]. Deep learning is a subset of machine learning designed to mimic the human brain through the use of artificial neural networks by using many layers and larger datasets. Generally, deep learning techniques are well suited for discovering and recognizing complex patterns in data that traditional machine learning methods can often miss. The increasing incorporation of deep learning in healthcare along with the availability of highly characterized cancer datasets has further accelerated research into the applications of deep learning in the analysis of the biology of cancer [27].

Given the complexities of diagnosing a tissue of origin from a cancer that has spread throughout the body, previous investigators have applied machine learning methods to determine tissue of origin for metastasized cancers [28] [29]. Longstanding techniques of microarrays and PCR have been utilized for generation of machine learning models for CUP detection, including support vector machines with 89% accuracy [30] and k-nearest neighbor algorithm (kNN) with 82% accuracy. [31 ,32]. LoCUP a tissue of origin classifier, was the first ML model using a multinomial logistic regression classifier with ridge penalties to incorporate tumor purity and reached a 95.8% accuracy[33]. Cup AI Dx, [28] used messenger RNA (mRNA) gene expression data from the TCGA dataset to train a network based on the popular inception model [30] to identify the tissue of origin, achieving an accuracy of 96.7% on a validation set of 354 TCGA metastatic samples. The TOD-CUP method [29] addressed the variation in mRNA platforms and used a gene expression rank-based majority vote algorithm to achieve an overall accuracy of 94%. Early work with microRNAs and non-deep learning ML algorithms showed 84% accuracy with kNN models [34] and binary decision trees at 85% [35] . However, investigation of deep learning ML models may improve on these accuracies with tissue of origin detection by microRNA. microRNAs are also at the forefront of extracellular vesicle liquid biopsy development and may be better suited for non-invasive classification of tissue of origin [36].

In this study, we set out to explore the possibility of developing a model for using microRNA profiles from metastatic tissues to determine the TOO through the application of deep-learning techniques. Successful TOO detection from microRNAs will provide a route for cancer detection without requiring samples from the primary tumor site in cases of CUP malignancies. We hypothesize that we would be able to predict the origin of metastatic tumors with a higher accuracy than previous reports by leveraging larger datasets of microRNA profiles from both normal and primary site tissues to train the model.

The data for this project was collected from the Cancer Genome Atlas (TCGA) dataset [6] and the SRA [7] from miTED. The TCGA dataset contains samples from 18 different cancer types representing 9648 samples, of which 365 were metastatic, 633 were solid normal, and 8650 were from the primary tumor site. Each sample consisted of microRNA expression data, available as RPM (reads per million mapped reads) as well as metadata including age and gender. We split the TCGA dataset into a combined primary tumor/solid normal samples training set and a metastatic sample test set. We then further split the primary tumor and solid normal samples into a training and validation set with a 9:1 ratio. The training set consisted of 8355 samples and the validation set consisted of 928 samples.

We use two datasets for evaluating the performance of our models. The SRA test dataset consisted of 194 samples from five different cancer types, all of which were from the primary tumor. We also use the metastatic samples from the TCGA dataset as our final test dataset, which contained samples from six cancer types. We developed four machine learning models, a decision tree classifier, random forest, logistic regression and finally a deep learning model. Our deep learning model performed with highest accuracy, achieving an accuracy of 97.0% on detecting TOO for metastatic samples and 80.4% on the non-metastatic SRA cohort. Feature importance analysis revealed the top three differentiating miRNA targets as mir-10b, mir-196b and mir-205, which confirms prior investigations on microRNAs associated with metastatic cancer [8,9,10].

METHODS

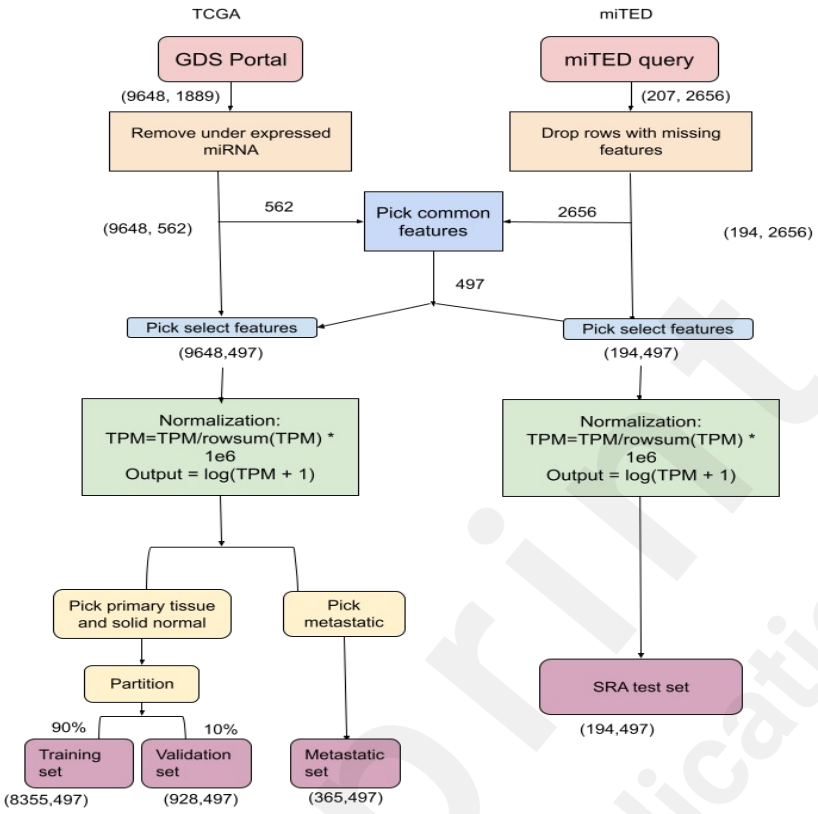


Figure 1: Overview of our data processing pipeline. *Data from the TCGA GDS portal and SRA miTED portal was obtained. Underexpressed microRNA and samples containing missing features from the miTED data were filtered. Common features were selected between both datasets, reducing the number of microRNA to 497. Features were normalized as RPM per sample and log transformed. The TCGA dataset was split into 1) the primary tissue and solid normal set and 2) the metastatic test set. The first, combined, set was further split into a training and validation set.*

In Figure 1, we outline the data preprocessing pipeline. Our study analyzed published data and did not generate any new sequencing data. TCGA data was obtained from [6]. Data was further filtered by querying the GDC via the APIs specified in [37]. We restricted the tissue type to be one of primary tumor, solid tissue normal, or metastatic. We further restricted the data as microRNA transcriptome profiling and picked data corresponding to 18 types of cancer each containing a sufficient number of samples, obtaining 9,648 files (Figure 2, Table S1).

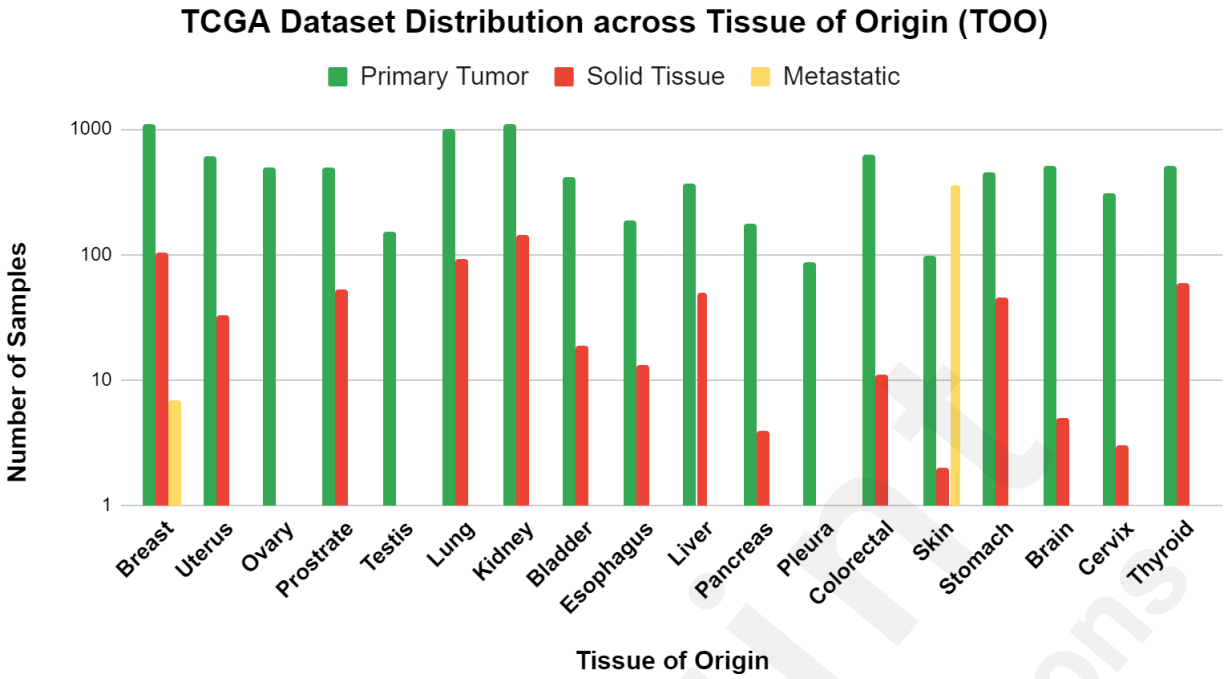


Figure 2: TCGA Dataset Distribution across TOO. *Distribution of the different cancer samples in the TCGA dataset that are from the primary tumor site, solid tissue, or metastatic. Note that metastatic samples primarily corresponded to skin as the tissue of origin.*

To obtain the SRA data, we used the miTED portal and restricted the cancer types to six types of cancer, seen in further detail in Figure 2. We obtained 207 samples, each containing expression data for 2656 microRNAs. After removing samples with missing features, 194 samples were remaining.

We selected miRNA features that were expressed in at least 50% of the samples, which reduced the number of features in the TCGA dataset from 1889 to 562. We then picked the common features between the SRA dataset and the TCGA dataset, reducing this number to 497. On both datasets, we normalized the RPM of the selected features per sample to sum to a million. We then transformed the RPM values using the transformation log(RPM + 1) to restrict the range of the input.

For implementation of decision tree, random forest, and logistic regression classifiers, the sklearn package was used [38]. We used classification accuracy as the primary metric to evaluate our models. Deep learning models were created with pytorch [39]. To optimize and train our neural network we utilized Adam optimizer and trained for 50 epochs. Since our objective was classification, we used softmax with cross entropy loss [40] to optimize the model. We used the validation set to determine the hyperparameters of the models, and picked the best performing model for further evaluation on the test set. Feature importance was calculated with sklearn's permutation feature importance function.

We share our implementation at [41]. The implementation script for selecting the specific queried data described above is available in gdc_query.py file in our repository. Code for implementing age and microRNA normalization is available in the file process_data.py. The implementation for model training, performance evaluation, and feature importance is available in the notebook miRNA_model_training_eval.ipynb.

RESULTS

In order to develop a model to detect TOO, we set out to find the best performing machine learning model for determining the TOO from the TCGA primary tumor and solid normal tissue cohorts. The models were then tested on the validation set, and we could accurately determine the TOO based on primary/normal miRNA profiles, with an accuracy of over 90% for 15 of 18 different tissue types using deep-learning (Figure 3 and Table S2).

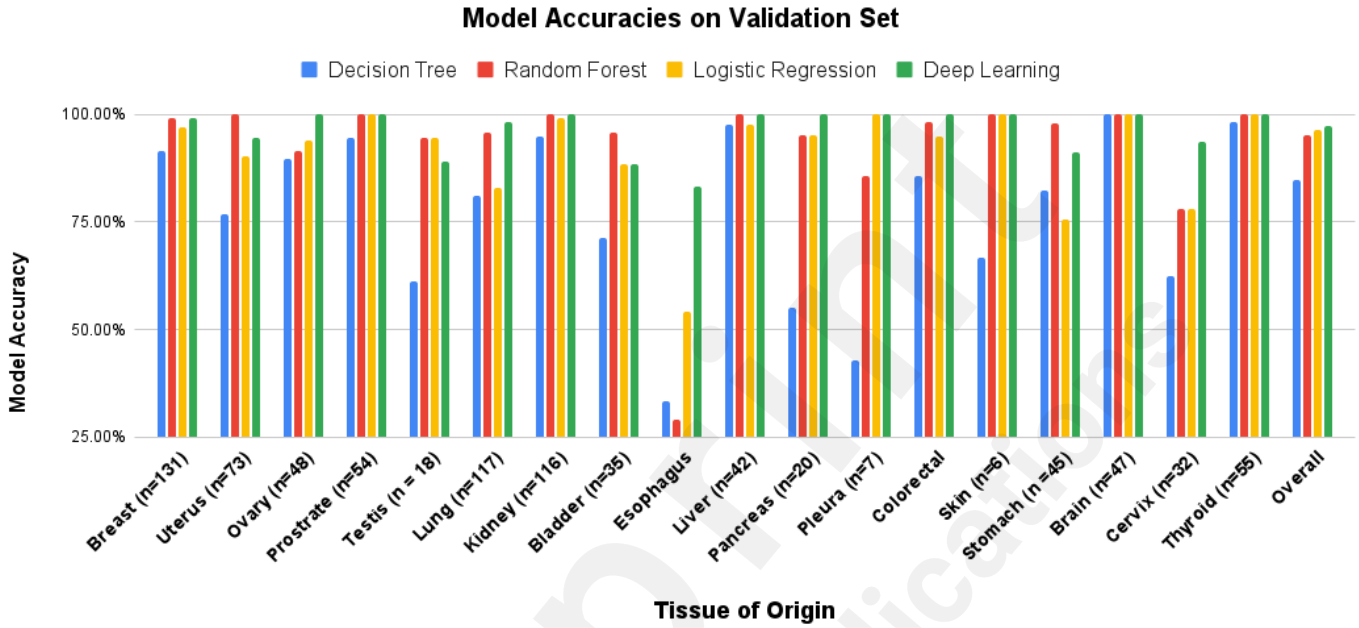


Figure 3: Model Accuracies on validation test set. *Performance of four models for the identification of tissue of origin. The validation set consists of both primary tumor and solid normal tissue samples from the TCGA dataset.*

We can note that the deep learning model performs consistently the highest on the validation set, with logistic regression and random forest classifiers providing comparable performance.

We then set out to apply our best performing deep learning model and evaluate its performance on the SRA test set that contains miRNA expression data from primary tumors (Table 1). We accurately determined the TOO with an accuracy of over 90% for three of the five cancer types but saw a decrease in accuracy for bladder and colorectal cancer.

Cancer Type	SRA Test Accuracy - Deep Learning
Breast (n = 44)	91.6%
Prostrate (n =37)	100%
Lung (n = 19)	100%
Bladder (n = 10)	80%
Colorectal (n=78)	58.9%
Skin (n=0)	N/A

Overall - Across Cancer Types	**80.4%**

Table 1: Performance of Deep Learning Model on SRA test set based on tissue type. *Performance of our deep learning model for the identification of tissue of origin on the primary tissue site cohorts from the SRA.*

Lastly, we analyzed our deep learning model on microRNA expression data from metastatic tissue samples in the TCGA dataset (Table 2). We accurately determined the TOO with an accuracy of over 85% for all cancer types with an average of 97%.

Cancer Type	TCGA Metastatic Test Accuracy - Deep Learning
Breast (n = 7)	85.7%
Prostrate (n = 1)	100%
Lung (n = 0)	N/A
Bladder (n = 1)	100%
Colorectal (n =1)	100%
Skin (n = 352)	97.4%
Overall - Across Cancer Types	**97.0%**

Table 2: Performance of Deep Learning Model on TCGA test set based on tissue type *Performance of our deep learning model for the identification of tissue of origin in metastatic tumor tissue.*

Since random forest and logistic regression classifiers provided comparable performance on the primary/normal validation set, we compared the classifier accuracy on both test sets for all created models (Table 3).

Classifier	Accuracy on TCGA metastatic test set	Accuracy on SRA test set
Decision Tree	62.5%	41.2%
Random Forest	94.2%	74.2%
Logistic Regression	93.2%	71.6%
Deep Learning	97.0%	80.4%

Table 3: Accuracy of developed models on metastatic and SRA test sets. *The accuracy for all four models is presented on the TCGA metastatic and SRA cohorts. The decision tree classifier had a depth of 14 and the random forest had a depth of 19.*

The input features of our models consist of microRNA expression data common to TCGA and SRA data sets. Figure 4 describes the overall architecture of the model which consists of 2 linear layers.

The second layer has 18 outputs, corresponding to each cancer type. The cancer type corresponds to the output with the maximum value.

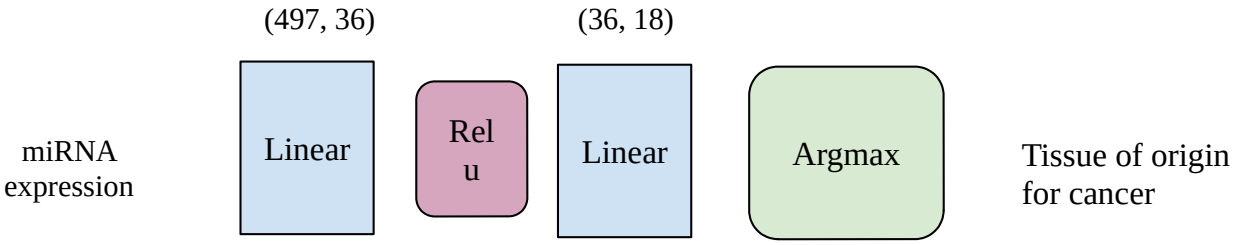


Figure 4: Deep Learning model architecture. *A schematic of the machine learning model architecture.*

We used dropout for the input layer [42] as it is a common technique to improve model accuracy and reduce overfitting. We also augmented our input data with noise.

To evaluate the performance of our models, we computed confusion matrices for performance on metastatic samples (Figure S2A,B) and plotted the ROC curves for performance on metastatic skin cancer (Figure S2C,D), as the majority of the metastatic samples were obtained from skin cancer cases. We observed that the deep learning model performed significantly better than our decision tree model, which was consistent when evaluated on the SRA validation cohort (Figure S3). To illustrate the effectiveness of our models, we created Sankey plots representing the deep learning model performance on metastatic samples from the TCGA dataset and primary tissue sites from the SRA dataset (Figure 5).

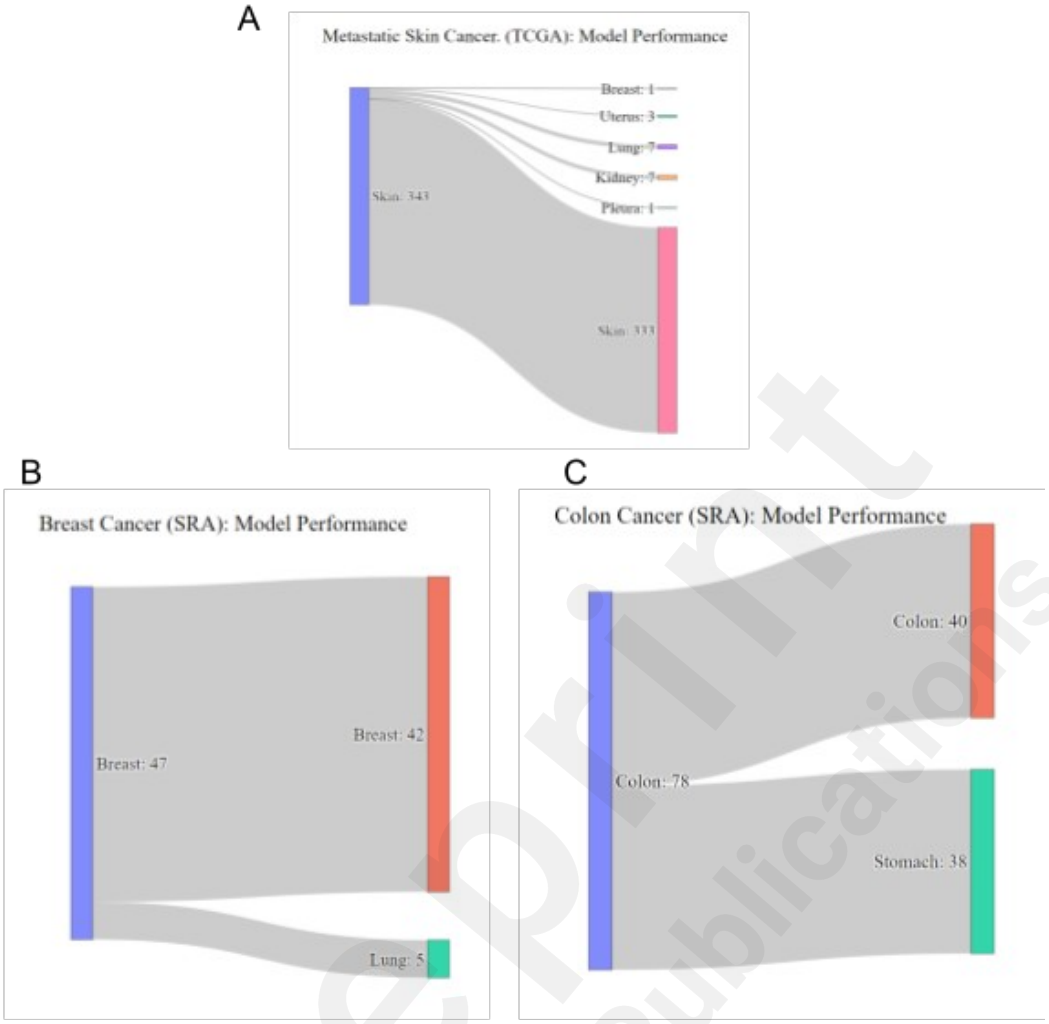


Figure 5: Sankey Plot for Deep Learning model on SRA and TCGA test datasets. *A) On the TCGA dataset, our deep learning model is able to correctly classify 333 out of 343 metastatic skin cancer samples, demonstrating high accuracy. B) On the SRA test dataset, we show representative plots for breast and colon cancers, showing high accuracy for breast cancer tissue of origin identification. The model performance on colon cancer is less accurate due to miRNA expression consistently overlapping for colon and stomach cancers [43]*

These results confirm our hypotheses and show that we were able to predict the TOO with high accuracy using deep learning. Furthermore, our findings demonstrated that deep learning techniques significantly increase the accuracy in comparison to decision tree, logistic regression and random forest models.

To reveal the significance of individual features, we performed feature importance analysis using permutation feature importance method (Figure 6A). The top three microRNAs contributing to our deep learning model based on our combined normal and primary site training set are mir-10b, mir-196, and mir-205. Mir-10b has been shown to function as a metastasis promoting factor in many cancer types. In fact, it was one of the first microRNAs to have been discovered with aberrant expression in cancer cells [8]. Mir-196 has been linked to the progression of many cancers, notably metastatic colorectal cancer [9] while mir-205 expression is downregulated in metastatic breast and prostate cancer [10].

To further understand the significance of the identified important features, we compute a heatmap

(Figure 6B) showing the miRNA expression values for the top ten miRNA features for samples in the training dataset. Visually, it is apparent that the miRNA features can be used to distinguish the cancer type. To further validate this, we perform PCA and t-SNE analysis using only the top ten features (Figure 6C, 6D). We note that the t-SNE plot shows clear separation of features into distinct clusters corresponding to each cancer type, showing the significance of the features for detecting the TOO.

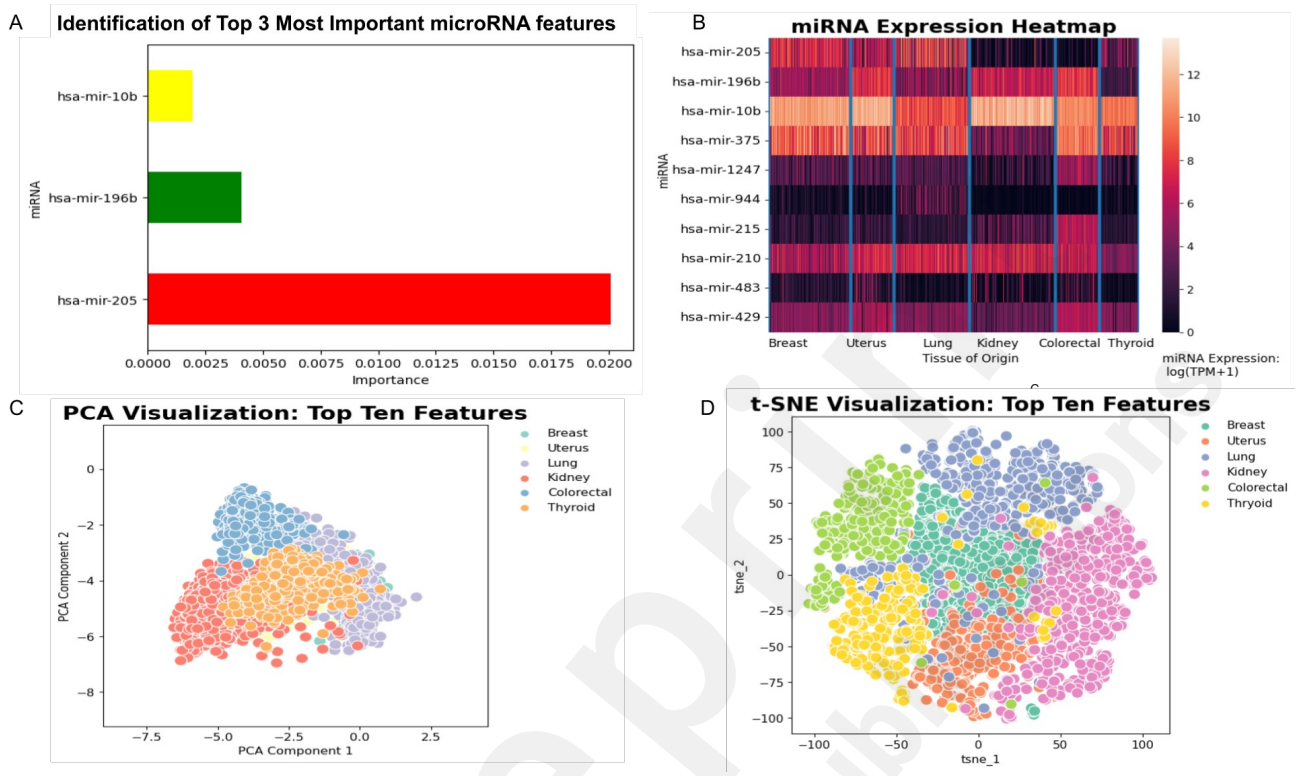


Figure 6: miRNA Feature Importance Visualizations *A) Permutation feature importance for the top three microRNA candidates. A bar graph of the importance values for the three top microRNA candidates for the Logistic Regression Model. B) miRNA Expression Heatmap. miRNA expression values for the top 10 most important features (as determined by permutation feature importance) for a subset of samples.The top 10 miRNA features can cluster cancer type. Low mir-205 and mir-944 and a high mir-10b is indicative of colorectal cancer. Similarly, low expressions for mir-429, mir-483, mir-215, mir-944, mir-1247, mir-375 and mir-205 are indicative of kidney cancer. C) PCA Visualization. D) TSNE Visualization. PCA and t-SNE visualization of data corresponding to the six cancer types with the most samples in our dataset, using only the top 10 miRNA features. In the PCA plot, note that there is significant overlap between the cancer types, while in the t-SNE plot, the cancer types are well separated. Suggesting that with ten miRNA features, machine learning models may correctly identify patterns and predict TOO.*

DISCUSSION

In these investigations, while employing successively more powerful classifiers, we were able to detect the TOO on solely metastatic cancer samples with accuracies ranging from 62.5% with a decision tree to 97.0% with a deep learning model. Our methods show that one can leverage larger amounts of gene expression data for primary and solid tissue normal tumor samples (~10000 samples) to come up with accurate classifiers to determine TOO for metastatic cancer (currently limited to ~300 samples). In order to verify the robustness of our model, we assessed its performance on primary tumor data from the SRA and obtained accuracies ranging from 41.2% with decision tree to 80.4% when employing deep learning. Our methods have also identified promising miRNA

candidates, reaffirming prior research in this field and demonstrating the potential of machine learning.

The predominant failure of our model on the SRA test cohort was within colorectal cancer as can be seen in Figure 5C. Many colorectal samples were incorrectly classified as stomach/gastric cancer. This is consistent with previous research in this area as miRNA expression profiles for gastro-intestinal cancers show significant overlap [43]. In addition, colorectal and stomach cancer are often synchronous with probabilities ranging from 20.1% - 37.2% [44].

We used permutation feature importance, a model agnostic metric that permutes features across samples in the test set to assess the change in model accuracy. The results are in line with existing research in this area and serve as a good indicator of the feasibility of machine learning techniques to identify promising biomarkers.

To effectively utilize our model in clinical care, accuracy must be improved further. Our model currently performs with an accuracy of 97.0%. While this may seem impressive, clinical classifiers should be highly accurate so that there are a negligible number of cases with errors in identifying TOO. To improve the accuracy, accumulation of larger datasets is necessary, and as the non-coding genome continues to reveal significant contributions to cancer, we predict that available datasets will expand. A further limitation to our study is that the available miRNA metastatic datasets are predominantly skin cancer. Thus, access to a larger, more varied, dataset would improve our assessment of model performance. Furthermore, in order to develop a truly noninvasive method of TOO identification relevant for all cancers, it would be ideal to extend our method to microRNA expression data from blood samples. Detecting the TOO through blood-based microRNA biomarkers would significantly impact the diagnosis and treatment of CUP patients. Additionally, our model cannot differentiate between tumor and solid tissue normal samples, as it was designed to identify the TOO specifically.

To summarize, our developed machine learning models can accurately identify the tissue of origin with high accuracy from microRNA expression data when trained on primary tumor and solid tissue samples. Importantly, our results identified key microRNA differentiators of tissue type. Our models are robust and perform well across different datasets (TCGA and the SRA dataset). We look forward to developing further deep learning models that can accurately detect TOO as microRNA datasets expand, with the goal of having a non-invasive test for diagnosing the presence of cancer and determining the cancer TOO with high accuracy.

ACKNOWLEDGMENTS
We are grateful to the Cancer Genome Atlas Project (TCGA) and patients for providing the data used in this research. We are thankful to Soroush Hajizadeh for reviewing our code and providing insightful feedback. The results are in part based upon data generated by the TCGA Research Network [6].

References:

[1] Qaseem A, Usman N, Jayaraj JS, Janapala RN, Kashif T. Cancer of unknown primary: A review on clinical guidelines in the development and targeted management of patients with the Unknown Primary Site. Cureus. Published online 2019. doi:10.7759/cureus.5552
[2] Massard C, Loriot Y, Fizazi K. Carcinomas of an unknown primary origin—diagnosis and treatment. Nature Reviews Clinical Oncology. 2011;8(12):701-710. doi:10.1038/nrclinonc.2011.158
[3] Bianchi JJ, Zhao X, Mays JC, Davoli T. Not all cancers are created equal: Tissue specificity in cancer genes and pathways. Current Opinion in Cell Biology. 2020;63:135-143.

doi:10.1016/j.ceb.2020.01.005
[4] Laprovitera N, Riefolo M, Ambrosini E, Klec C, Pichler M, Ferracin M. Cancer of Unknown Primary: Challenges and Progress in Clinical Management. Cancers (Basel). 2021;13(3):451. Published 2021 Jan 25. doi:10.3390/cancers13030451
[5] Chen B, Dragomir MP, Yang C, Li Q, Horst D, Calin GA. Targeting non-coding RNAs to overcome cancer therapy resistance. Signal Transduct Target Ther. 2022;7(1):121. Published 2022 Apr 13. doi:10.1038/s41392-022-00975-3
[6] The Cancer Genome Atlas Program (TCGA)- National Cancer Institute. https://www.cancer.gov/ccg/research/genome-sequencing/tcga.
[7] Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011;39(Database issue):D19-D21. doi:10.1093/nar/gkq1019
[8] Sheedy P, Medarova Z. The fundamental role of miR-10b in metastatic cancer. Am J Cancer Res. 2018;8(9):1674-1688. Published 2018 Sep 1.
[9] Chauhan N, Dhasmana A, Jaggi M, Chauhan SC, Yallapu MM. miR-205: A Potential Biomedicine for Cancer Therapy. Cells. 2020;9(9):1957. Published 2020 Aug 25. doi:10.3390/cells9091957
[10] Lu YC, Chang JT, Chan EC, Chao YK, Yeh TS, Chen JS, Cheng AJ. miR-196, an Emerging Cancer Biomarker for Digestive Tract Cancers. J Cancer. 2016 Mar 20;7(6):650-5. doi: 10.7150/jca.13460. PMID: 27076845; PMCID: PMC4829550.
[11] Chiam K, Mayne GC, Wang T, et al. Serum outperforms plasma in small extracellular vesicle microRNA biomarker studies of adenocarcinoma of the esophagus. World J Gastroenterol. 2020;26(20):2570-2583. doi:10.3748/wjg.v26.i20.2570
[12] Monzon FA, Medeiros F, Lyons-Weiler M, Henner WD. Identification of tissue of origin in carcinoma of unknown primary with a microarray-based gene expression test. Diagn Pathol. 2010;5:3. Published 2010 Jan 13. doi:10.1186/1746-1596-5-3
[13] Pu X, Yang S, Xu Y, Chen B, Wang Q, Gong Q and Wu L (2021) Case Report: Tissue Origin Identification for Cancer of Unknown Primary: Gene Expression Profiling Approach. Front. Oncol. 11:702887. doi: 10.3389/fonc.2021.702887
[14] Wei Tang, Shixiang Wan, Zhen Yang, Andrew E Teschendorff, Quan Zou, Tumor origin detection with tissue-specific miRNA and DNA methylation markers, Bioinformatics, Volume 34, Issue 3, February 2018, Pages 398–406, https://doi.org/10.1093/bioinformatics/btx622
[15] Greco, F.A., Burris, H.A., III, Erland, J.B., Gray, J.R., Kalman, L.A., Schreeder, M.T. and Hainsworth, J.D. (2000), Carcinoma of unknown primary site. Cancer, 89: 2655-2660. https://doi.org/10.1002/1097-0142(20001215)89:12<2655::AID-CNCR19>3.0.CO;2-9
[16] Schapira DV, Jarrett AR. The need to consider survival, outcome, and expense when evaluating and treating patients with unknown primary carcinoma. Arch Intern Med. 1995;155(19):2050-2054.
[17] Zhang B, Pan X, Cobb GP, Anderson TA. microRNAs as oncogenes and tumor suppressors. Dev Biol. 2007;302(1):1-12. doi:10.1016/j.ydbio.2006.08.028
[18] Smolarz, B., Durczyński, A., Romanowicz, H., Szyłło, K., & Hogendorf, P. (2022). miRNAs in Cancer (Review of Literature). *International journal of molecular sciences*, *23*(5), 2805. https://doi.org/10.3390/ijms23052805
[19] Chakrabortty, A., Patton, D. J., Smith, B. F., & Agarwal, P. (2023). miRNAs: Potential as Biomarkers and Therapeutic Targets for Cancer. *Genes*, *14*(7), 1375. https://doi.org/10.3390/genes14071375
[20] Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R., & Golub, T. R. (2005). MicroRNA expression profiles classify human cancers. *Nature*, *435*(7043), 834–838. https://doi.org/10.1038/nature03702
[21] Rosenfeld N, Aharonov R, Meiri E, et al. MicroRNAs accurately identify cancer tissue origin.

Nat Biotechnol. 2008;26(4):462-469. doi:10.1038/nbt1392
[22] Tucci P. The Role of microRNAs in Cancer: Functions, Biomarkers and Therapeutics. Cancers (Basel). 2022 Feb 10;14(4):872. doi: 10.3390/cancers14040872. PMID: 35205620; PMCID: PMC8870119.
[23] Zhang, B., Shi, H., & Wang, H. (2023). Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. *Journal of multidisciplinary healthcare*, *16*, 1779–1791. https://doi.org/10.2147/JMDH.S410301
[24] Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
[25] Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
[26] Chung, M. K. (2020). Introduction to logistic regression. *arXiv preprint arXiv:2008.13567*.
[27] Tran, K.A., Kondrashova, O., Bradley, A. *et al.* Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* 13, 152 (2021). https://doi.org/10.1186/s13073-021-00968-x
[28] Zhao Y, Pan Z, Namburi S, et al. Cup-ai-DX: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine. 2020;61:103030. doi:10.1016/j.ebiom.2020.103030
[29] Shen, Y., Chu, Q., Yin, X., He, Y., Bai, P., Wang, Y., Fang, W., Timko, M. P., Fan, L., & Jiang, W. (2021). TOD-CUP: a gene expression rank-based majority vote algorithm for tissue origin diagnosis of cancers of unknown primary. *Briefings in bioinformatics*, *22*(2), 2106–2118. https://doi.org/10.1093/bib/bbaa031
[30] Tothill, R. W., Kowalczyk, A., Rischin, D., Bousioutas, A., Haviv, I., van Laar, R. K., ... & Holloway, A. J. (2005). An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Research*, *65*(10), 4031-4040.
[31] van Laar, R.K., Ma, X.-J., de Jong, D., Wehkamp, D., Floore, A.N., Warmoes, M.O., Simon, I., Wang, W., Erlander, M., van't Veer, L.J. and Glas, A.M. (2009), Implementation of a novel microarray-based diagnostic test for cancer of unknown primary. Int. J. Cancer, 125: 1390-1397. https://doi.org/10.1002/ijc.24504
[32] Xiao-Jun Ma, Rajesh Patel, Xianqun Wang, Ranelle Salunga, Jaji Murage, Rupal Desai, J. Todd Tuggle, Wei Wang, Shirley Chu, Kimberly Stecker, Rajiv Raja, Howard Robin, Mat Moore, David Baunoch, Dennis Sgroi, Mark Erlander; Molecular Classification of Human Cancers Using a 92-Gene Real-Time Quantitative Polymerase Chain Reaction Assay. *Arch Pathol Lab Med* 1 April 2006; 130 (4): 465–473. doi: https://doi.org/10.5858/2006-130-465-MCOHCU
[33] Søndergaard, D., Nielsen, S., Pedersen, C. & Besenbacher, S. (2017). Prediction of Primary Tumors in Cancers of Unknown Primary. *Journal of Integrative Bioinformatics*, *14*(2), 20170013. https://doi.org/10.1515/jib-2017-0013
[34 Varadhachary, G. R., Spector, Y., Abbruzzese, J. L., Rosenwald, S., Wang, H., Aharonov, R., ... & Raber, M. N. (2011). Prospective gene signature study using microRNA to identify the tissue of origin in patients with carcinoma of unknown primary. *Clinical Cancer Research*, *17*(12), 4063-4070.
[35] Rosenwald, S., Gilad, S., Benjamin, S., Lebanony, D., Dromi, N., Faerman, A., ... & Aharonov, R. (2010). Validation of a microRNA-based qRT-PCR test for accurate identification of tumor tissue origin. *Modern Pathology*, *23*(6), 814-823.
[36] de Miguel Pérez, D., Rodriguez Martínez, A., Ortigosa Palomo, A. *et al.* Extracellular vesicle-miRNAs as liquid biopsy biomarkers for disease identification and prognosis in metastatic colorectal cancer patients. *Sci Rep* 10, 3974 (2020). https://doi.org/10.1038/s41598-020-60212-1
[37] GDC Application Programming Interface (API). NCI Genomic Data Commons. https://gdc.cancer.gov/developers/gdc-application-programming-interface-api.
[38] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[39] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. arXiv.org, 2019. https://doi.org/10.48550/arXiv.1912.01703.
[40] Mao, A., Mohri, M., & Zhong, Y. (2023, July). Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning* (pp. 23803-23828). PMLR.
[41] Anisha234. ANIHA234/identifying-tissue-of-origin-from-mirna. GitHub. 2023. https://github.com/Anisha234/miRNA.
[42] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. Published online 2014. doi:https://doi.org/10.48550/arXiv.1207.0580
[43] Laprovitera, N., Riefolo, M., Porcellini, E., Durante, G., Garajova, I., Vasuri, F., Aigelsreiter, A., Dandachi, N., Benvenuto, G., Agostinis, F., Sabbioni, S., Berindan Neagoe, I., Romualdi, C., Ardizzoni, A., Trerè, D., Pichler, M., D'Errico, A., & Ferracin, M. (2021). MicroRNA expression profiling with a droplet digital PCR assay enables molecular diagnosis and prognosis of cancers of unknown primary. Molecular oncology, 15(10), 2732–2751. https://doi.org/10.1002/1878-0261.13026
[44] Suh B. J. (2016). Synchronous and Metachronous Colon Cancers in Patients with Gastric Cancer: Report of 2 Cases. *Case reports in oncology*, *9*(3), 752–759. https://doi.org/10.1159/000452831

Supplementary Files

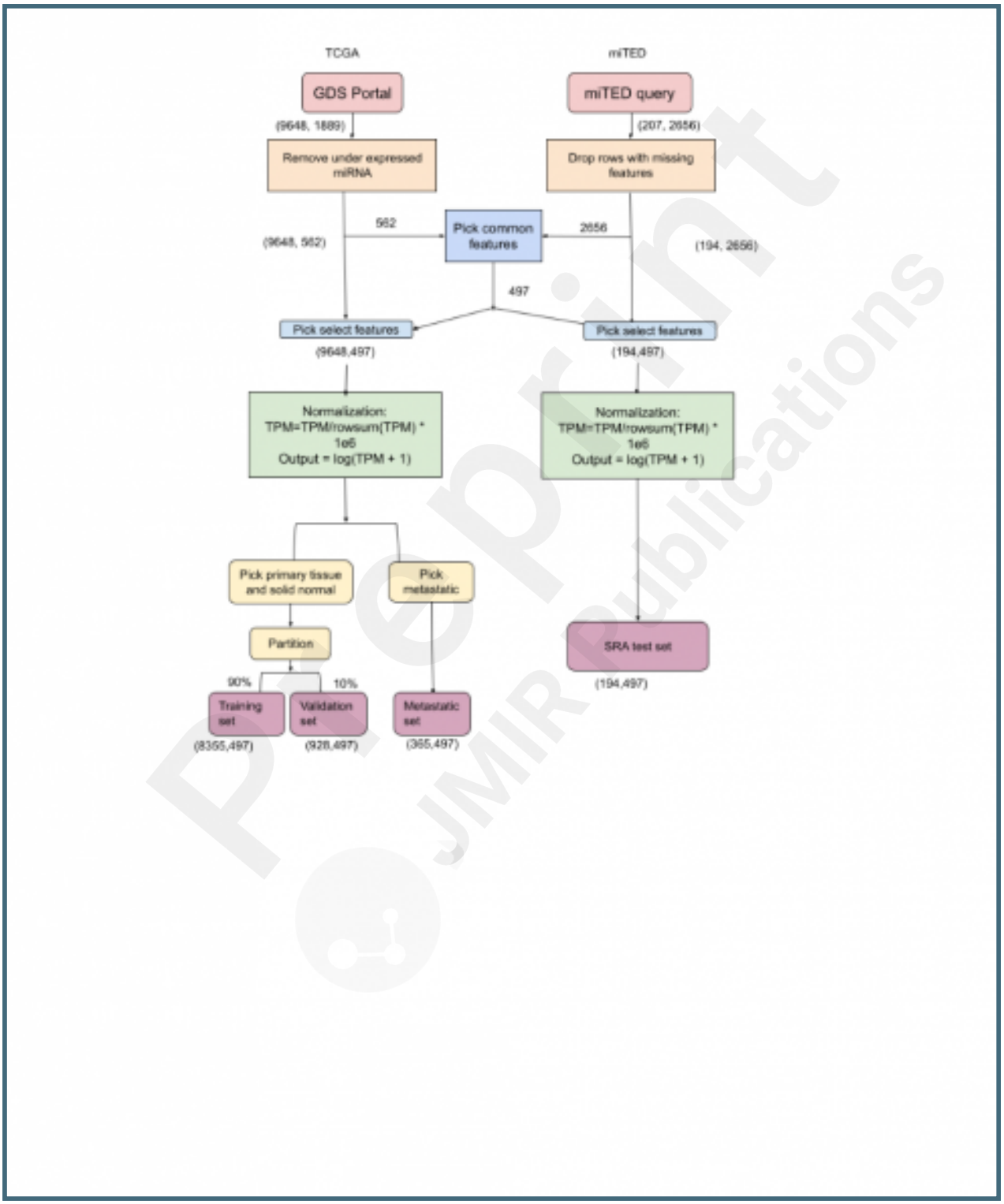
Untitled.
URL: http://asset.jmir.pub/assets/54dd5b031f138b18c76334ca812425ab.docx

Untitled.
URL: http://asset.jmir.pub/assets/5fb8298ca96d5f291440ac52d3582249.pdf

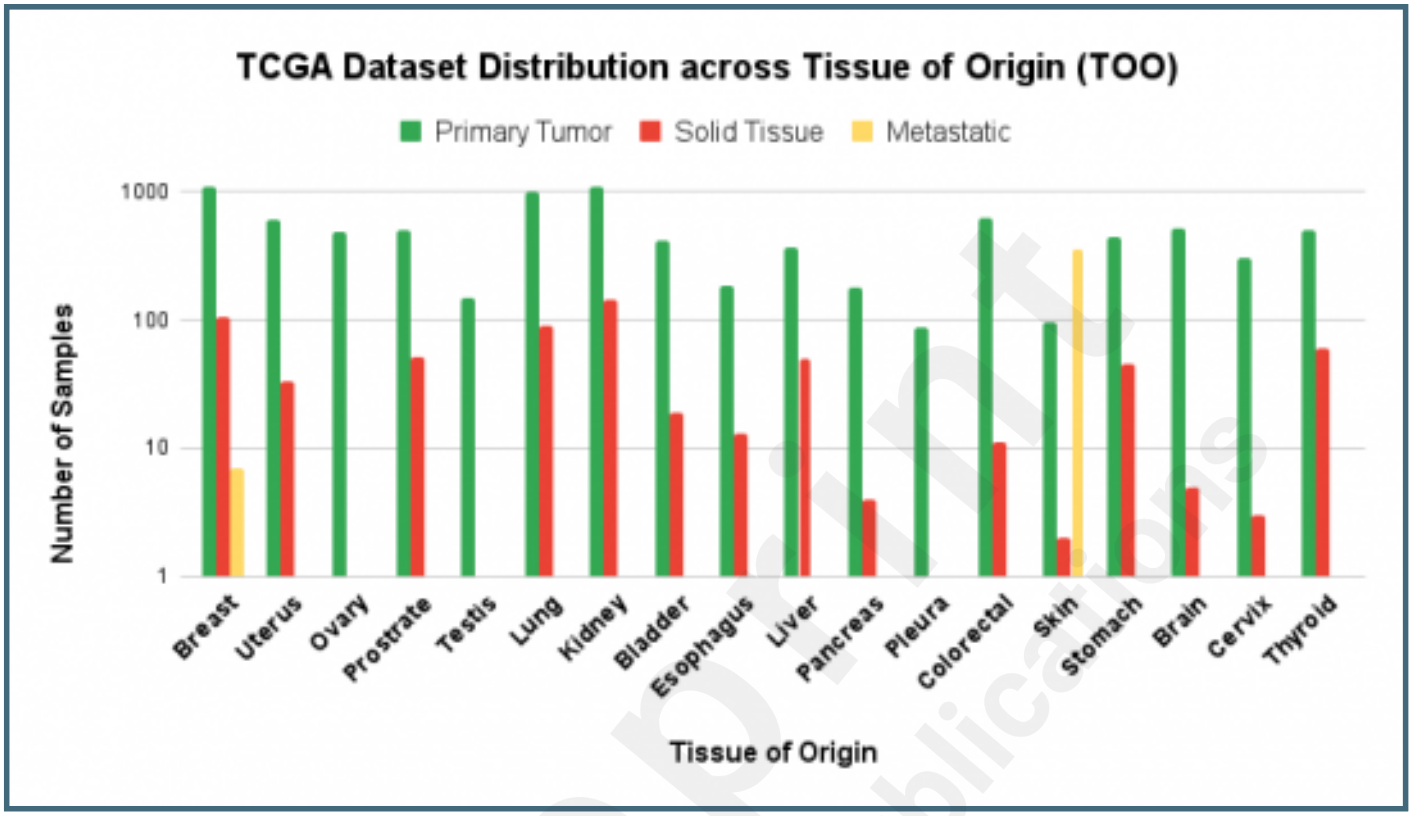
Untitled.
URL: http://asset.jmir.pub/assets/d7e054e62b00b5f87ec2e87b41ee4d16.docx

Figures

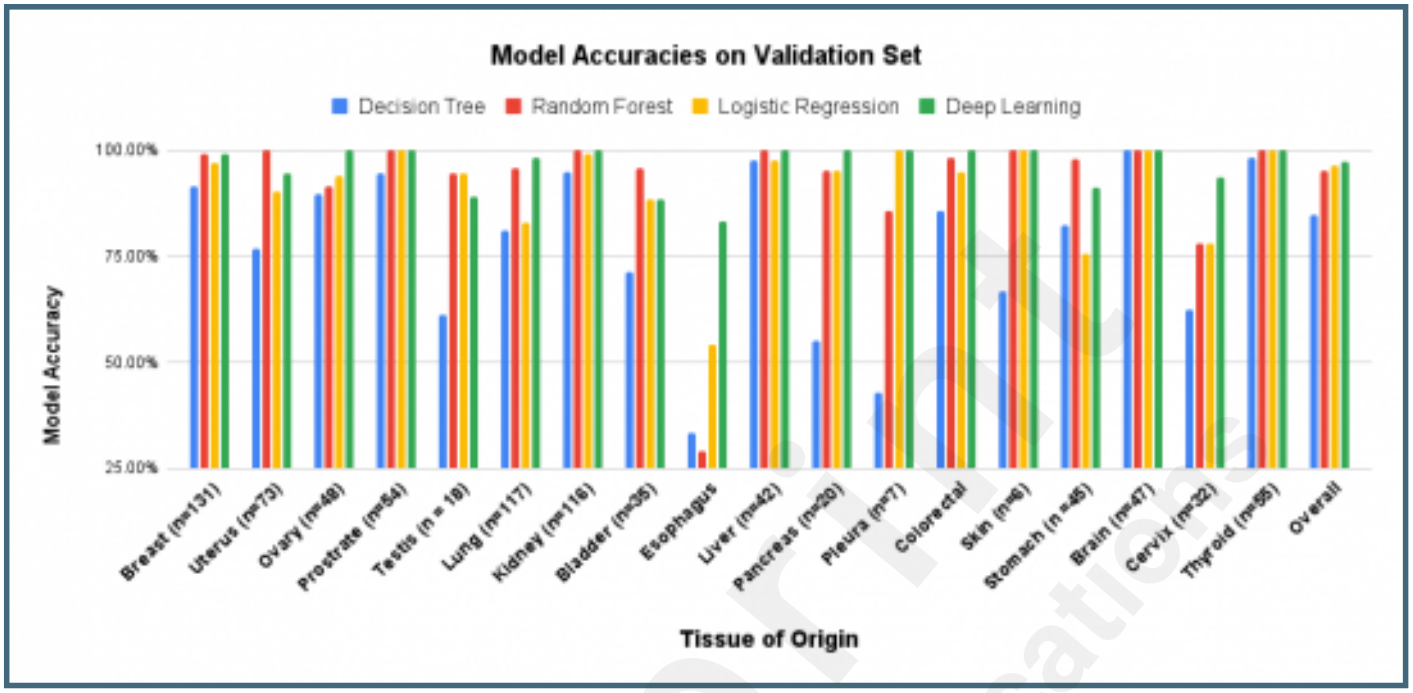
Overview of our data processing pipeline. Data from the TCGA GDS portal and SRA miTED portal was obtained. Underexpressed microRNA and samples containing missing features from the miTED data were filtered. Common features were selected between both datasets, reducing the number of microRNA to 497. Features were normalized as RPM per sample and log transformed. The TCGA dataset was split into 1) the primary tissue and solid normal set and 2) the metastatic test set. The first, combined, set was further split into a training and validation set.



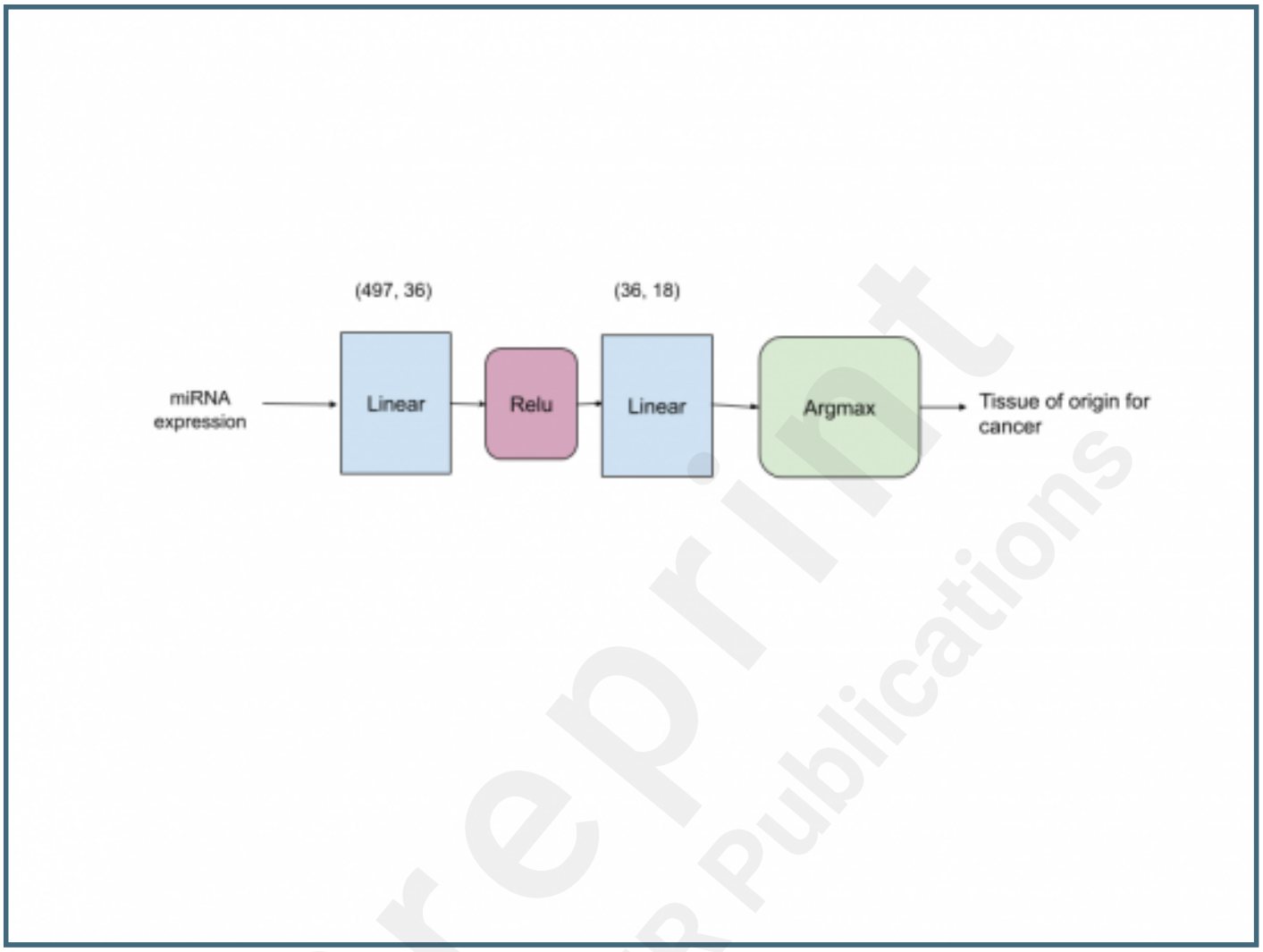
TCGA Dataset Distribution across TOO. Distribution of the different cancer samples in the TCGA dataset that are from the primary tumor site, solid tissue, or metastatic. Note that metastatic samples primarily corresponded to skin as the tissue of origin.



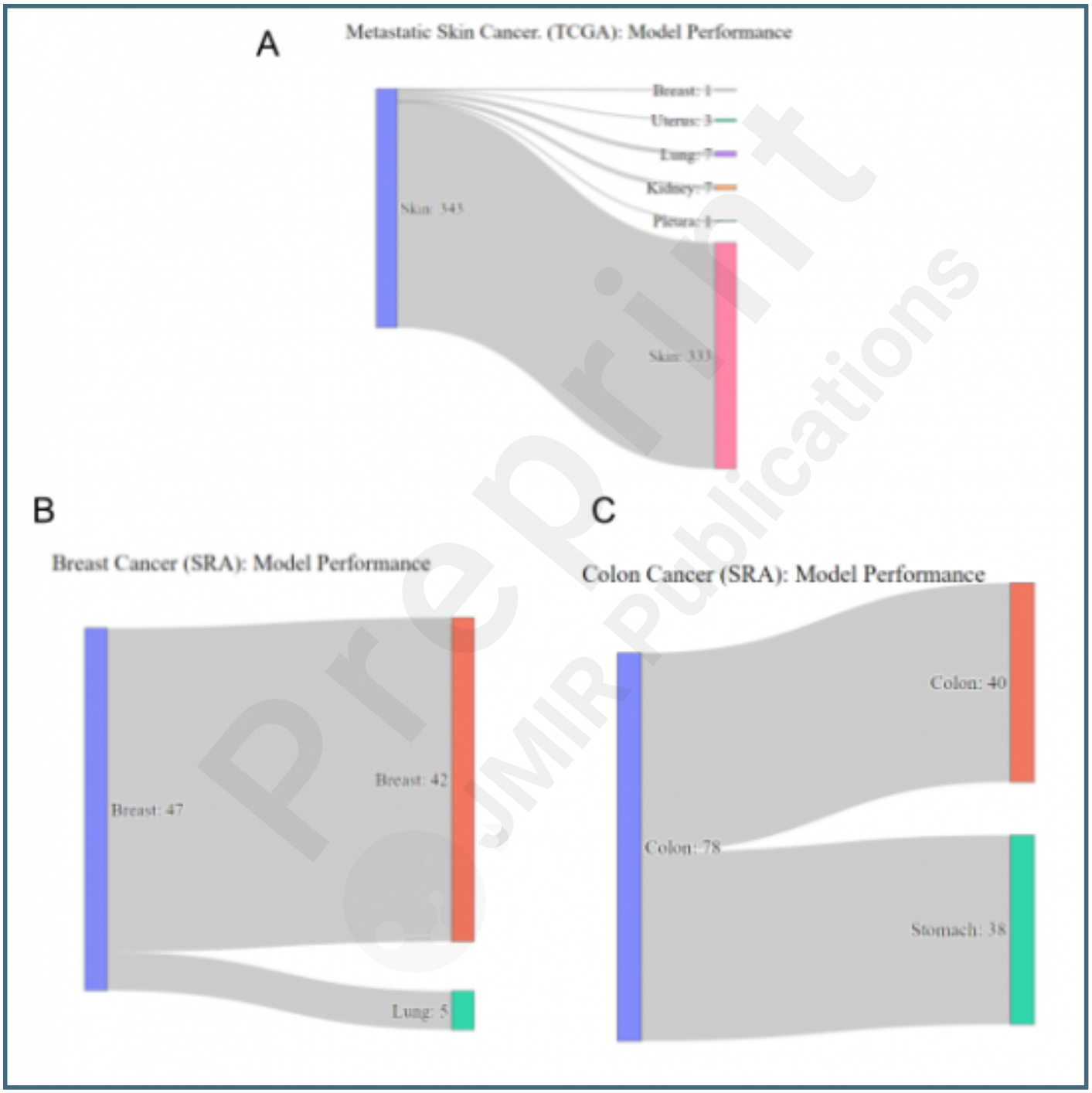
Model Accuracies on validation test set. Performance of four models for the identification of tissue of origin. The validation set consists of both primary tumor and solid normal tissue samples from the TCGA dataset.



Deep Learning model architecture. A schematic of the machine learning model architecture.



Sankey Plot for Deep Learning model on SRA and TCGA test datasets. A) On the TCGA dataset, our deep learning model is able to correctly classify 333 out of 343 metastatic skin cancer samples, demonstrating high accuracy. B) On the SRA test dataset, we show representative plots for breast and colon cancers, showing high accuracy for breast cancer tissue of origin identification. The model performance on colon cancer is less accurate due to miRNA expression consistently overlapping for colon and stomach cancers [43].



miRNA Feature Importance Visualizations A) Permutation feature importance for the top three microRNA candidates. A bar graph of the importance values for the three top microRNA candidates for the Logistic Regression Model. B) miRNA Expression Heatmap. miRNA expression values for the top 10 most important features (as determined by permutation feature importance) for a subset of samples.The top 10 miRNA features can cluster cancer type. Low mir-205 and mir-944 and a high mir-10b is indicative of colorectal cancer. Similarly, low expressions for mir-429, mir-483, mir-215, mir-944, mir-1247, mir-375 and mir-205 are indicative of kidney cancer. C) PCA Visualization. D) TSNE Visualization. PCA and t-SNE visualization of data corresponding to the six cancer types with the most samples in our dataset, using only the top 10 miRNA features. In the PCA plot, note that there is significant overlap between the cancer types, while in the t-SNE plot, the cancer types are well separated. Suggesting that with ten miRNA features, machine learning models may correctly identify patterns and predict TOO.