Article

Biomarker Discovery for Meta-Classification of Melanoma Metastatic Progression using Transfer Learning

Jose Marie Antonio Minoza [1,*], Jonathan Adam Rico [2], Pia Regina Fatima Zamora [2], Manny Bacolod [3], Reinhard Laubenbacher [4,*], and Romulo de Castro [2,5,*]

[1] System Modeling and Simulation Laboratory, Department of Computer Science, University of the Philippines Diliman, Quezon City, Philippines
[2] Center for Informatics, University of San Agustin, Iloilo City, Philippines
[3] Department of Microbiology and Immunology, Weill Cornell Medicine, New York, United States
[4] Department of Medicine, University of Florida, Gainesville, Florida, United States
[5] 3R Biosystems, Long Beach, CA, United States
* Correspondence: jminoza@up.edu.ph, Reinhard.Laubenbacher@medicine.ufl.edu, rdecastro@usa.edu.ph

Simple Summary: The objective of this study is to leverage the use of machine learning techniques to inform on the process of metastatic progression in melanoma. The transfer-learning based biomarker discovery method we developed found known and novel biomarker genes for melanoma classification and survival prediction. Assessment of data representativeness allowed us to partially predict the patient population segments that the model will work in.

Abstract: Melanoma is considered the most serious and aggressive type of skin cancer, and metastasis appears to be the most important factor in prognosticating this type of cancer. With the emergence of new therapeutic strategies for metastatic melanoma that have shown improvement in patient survival, we developed a transfer learning-based biomarker discovery model that could help in the diagnosis and prognosis of this disease. After applying it to the ensemble machine learning model, results reveal that the genes we found show consistency with other methodologies previously applied to the same TCGA (The Cancer Genome Atlas) data set, and our methods found novel biomarker genes as well. Our ensemble model achieved Area Under the Receiver Operating Characteristic (AUC) of 0.9861, an accuracy of 91.05, and an F1 score of 90.60 using an independent validation data set. This study was able to identify potential genes for diagnostic classification (C7 and GRIK5) and diagnostic and prognostic biomarkers (S100A7, S100A7, KRT14, KRT17, KRT6B, KRTDAP, SERPINB4, TSHR, PVRL4, WFDC5, IL20RB). We also assessed the potential sources of bias for our model and confirmed some of them by the model's performance.

Keywords: melanoma; biomarker; transfer learning; ensemble model; bias; machine learning;

1. Introduction

Melanoma is a cancer that arises from pigment-containing cells called melanocytes. It is considered the most serious and aggressive type of skin cancer. [1,2] Its etiology is influenced by both genetics and environmental factors. [3–5]. Before being diagnosed, melanoma has often spread to a distant location [6]; and the majority of melanoma deaths are caused by these metastases.

Metastases appear to be the most significant factor influencing melanoma patients' prognosis. The advancement of new therapeutic strategies to extend patients' overall survival will benefit from research into the mechanisms of melanoma metastasis. Since the advent of new therapies and interventions, such as immune checkpoint inhibitors and targeted therapies for metastatic melanoma, mortality rates for melanoma have decreased by 6.4 percent per year in the United States from 2013 to 2017. [7,8]. To support these new treatments, novel molecular signatures that can be used for diagnosis, prognosis and treatment selection are needed. Such biomarkers may further reveal molecular mechanisms of melanoma metastasis that could help inform and improve patients' overall survival.

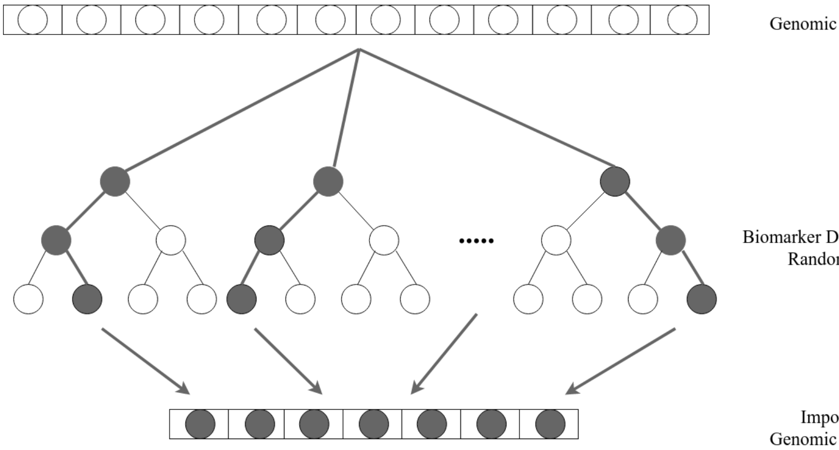
Gene expression profiling has been a powerful tool for identifying molecules involved in melanoma metastasis [9,10]. Machine learning techniques are considered a promising method in cancer prognostic development as genomic data have become more accessible. Despite this, it is extremely challenging given the high dimensionality of the data and the small number of patient samples. Several machine learning techniques have already been used as a disease classifier in melanoma but mostly focus on images [11–13], not yet genomic signatures which may be more informative and accurate. Compared to other related investigations [1,2,14,15], this study proposes transfer learning as a biomarker discovery technique and ensembles different classifiers that operate on different identified genomic signature subsets by soft voting. In addition, the level of expression of the weighted genomic biomarkers was investigated, in terms of survival of the patients, in order to gain a better understanding of melanoma metastasis and to identify potential therapeutic targets.

Lastly, preliminary data assessments allowed us to make predictions regarding bias and model performance, to better identify what subsets of patients the models could be applied to.

2. Materials and Methods

2.1. Transfer Learning for Biomarker Discovery

Machine learning (ML) algorithms that were developed to store the information acquired and applied to a different but related problem are referred to as transfer learning [16]. A large number of data and computing resources may be required to train a model; but transfer learning can possibly help address this issue. As a result, using transfer learning for datasets with high dimensionality and potentially complex interactions would be beneficial.



Discovery using RF.png

Figure 1. Biomarker Discovery using Random Forest

Biomarker discovery seeks to identify a subset of measured variables (biomarkers, whether they are genomic or clinical characteristics) that can be used to reliably predict a disease phenotype [17]. One of the popular approach in transfer learning is feature extraction; which in this case, it is to extract genomic features possibly responsible in melanoma progression. Rule-based transfer learning for biomarker discovery show improvement of classification performance, however, it has noticable poor performance on structure learning [17]. Random Forest (RF), on the other hand, appears to be good at finding interesting features in high-dimensional phenotype data with small key effects and low heritability. [18] This may be due to the way it accounts for potential gene-gene interactions when calculating significance scores for specific attributes, see Figure 1.

2.2. Protein-Protein Interaction Network

The complex interactions of all molecules describe biological processes best and determine various cellular functions and responses. Its mapping is a crucial step in (1) trying to unravel their unique molecular relationships in specific biological contexts and eventually (2) targeting of therapy for treatment of diseases, such as cancer. [19–21] Protein-protein interaction (PPI) networks are typically represented as graphs, with nodes representing proteins and edges connecting pairs of interacting proteins that are undirected and presumably weighted. [22].

Biomarker discovery using transfer learning could inform us of significant genomic features through computational methods; however, it is also important to get the nuances of biomarkers' roles and interactivity with other genes. Thus, identified genomic signatures as potential biomarkers were mapped into the whole network and the PPI network was then acquired. The PPI information used in this study was downloaded from STRINGDB (https://string-db.org/api/tsv/network), a database containing protein interactions that include physical and functional associations [23]. To identify which genes hold the most information, betweenness centrality is used, then the genes are ranked:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where v is the node gene retrieved from STRINGDB, σ_{st} is the total number of shortest paths from node gene s to node gene t, and $\sigma_{st}(v)$ is is the number of those paths that pass through v.

2.3. Clinical and Genomic Data

RNAseq and clinical data for skin cutaneous melanoma (SKCM) was retrieved from The Cancer Genome Atlas (TCGA) using the TCGAbiolinks [24,25] R package. The data set contains 365 metastatic and 103 primary tumor samples. Then, the normalized reads count (per million reads mapped) of RNA-seq underwent $\log_2$ transformation (all values less than 1 were assigned to 1 before transformation); we then carried out normalization of the data, since the level of expression of genes varied in different scales. To reduce low variance features, 0.95 was set as variance threshold; this led to the decrease in genomic features from 19947 to 19815 for training the machine learning model.

In this study, the underlying bias within the data set was assessed to ensure that the end users of the models are aware of the potential shortcomings when applied in the clinical setting (once validated).

2.3.1. Machine Learning Models

Machine learning techniques have been used in a variety of medical applications. However, it commonly used on imaging data such as ultrasound, xrays and slide specimens. [26–28]. Likewise in melanoma [29,30], computer vision is naturally to be used since it is first suspected visually through skin lesions. But according to a study [6], by the time melanoma is discovered, it has already metastasized. This study attempts to develop a meta-classification model that can determine late stage (metastatic) from early stage (primary tumor) melanoma using genomic data.

The biomarkers from both Random Forest, through feature importance scores, and PPI network, through betweenness centrality (BC) scores, were rank selected and applied to (i) Logistic Regression, (ii) Support Vector Machines, (iii) Gaussian Naive Bayes, and (iv) Random Forest. In classification models like those used for identifying the melanoma stage, the Area Under the Receiver Operating Characteristic (ROC) Curve, known as AUROC or AUC, gives the probability that a randomly selected melanoma patient with metastatic stage will have a higher predicted probability of being metastatic than a randomly selected melanoma patient with primary tumor stage.

DeLong's method [31] compares the performance of two models and accounts for the uncertainty caused by the finite training set's randomness and the evaluation on a common validation/test set. To calculate the z-score when comparing model A and model B in terms of AUC, the following is used:

$$z \triangleq \frac{\hat{\theta}^{(A)} - \hat{\theta}^{(B)}}{\sqrt{\mathbb{V}[\hat{\theta}^{(A)} - \hat{\theta}^{(A)}]}} = \frac{\hat{\theta}^{(A)} - \hat{\theta}^{(B)}}{\sqrt{\mathbb{V}[\hat{\theta}^{(A)}] + \mathbb{V}[\hat{\theta}^{(B)}] - 2\mathbb{C}[\hat{\theta}^{(A)}, \hat{\theta}^{(B)}]}} \tag{2}$$

where $\hat{\theta}^{(A)}, \hat{\theta}^{(B)}$ are AUC scores of the model A and B respectively, $\mathbb{V}$ is the variance and $\mathbb{C}$ is the covariance function. Under the null hypothesis [32], z can be well approximated by the standard normal distribution. Therefore, if the value of z deviates too much from zero, e.g., $z > 1.96$, it is, as a result, rational to consider that $\hat{\theta}^{(A)} > \hat{\theta}^{(B)}$ with the significance level $p < 0.05$. To put it another way, if z deviates too much from zero, we can infer that Model A has a statistically different AUC from Model B with $p < 0.05$.

Instead of committing completely to a single best classifier, two or more models that appear to complement each other, such as models that perform exceptionally well in different regions of the ROC space, could be combined. Hence, the models will be selected based on significant AUC scores, and are ensembled via soft voting.

In soft voting [33], the predicted class labels $\hat{y}$ based on the predicted probabilities p for each classifier is given by,

$$\hat{y} = argmax_i \sum_{j=1}^{m} w_i p_{ij} \tag{3}$$

where $i \in \{0, 1\}$ class labels and w_j is the weight that can be assigned to the jth classifier. In this study, weights were uniform across the classifier models.

2.3.2. Survival Analysis

Cancer prognosis deals with the probability of a patient surviving (and for how long) after being diagnosed with cancer. A commonly used tool [34] for modeling and visualizing patient survival is Kaplan-Meier analysis [35]. In the context of melanoma, the Kaplan-Meier curve describes the survival rate or the number of melanoma patients surviving at each time point from diagnosis, given by the survival function,

$$\hat{S}(t) = \prod_{t_i < t} (\frac{n_i - d_i}{n_i}) \tag{4}$$

where t is the elapsed time after diagnosis, d is the number of death events at time t, and n is the number of melanoma patients at risk at time t.

Davidson-Pilon's Lifelines KaplanMeierFitter (KMF) [36] Python module was used to estimate the survival function in Equation 4 and plot the survival curves. The KMF module requires two inputs, event E and duration T for which the patient was observed for event E. We used the *vital_status* field from TCGA as event E so that a value of one (1) means that death was observed while a value of zero (0) means right-censored (loss to follow up). For input T, we created another field, *days_to_event*, which is a combination of the *days_to_death* and *days_to_last_follow_up* fields of the TCGA data set such that the empty values in the *days_to_death* field is filled with *days_to_last_follow_up*.

Two Kaplan-Meier curves can be plotted on the same graph to determine if a certain variable (e.g. age, gender) produces statistically different survival rates. In this study, we aim to determine if certain genes (variable of interest) affect the prognosis of melanoma patients. That is, if a patient with high expression of a certain gene would yield to poor survivability or if a patient with low expression of a certain gene would yield to better survivability. After normalizing the data using the standard scaling per gene, we used the statistical mean as the threshold for high and low gene expressions. Log-rank test with $\alpha = 0.99$ indicates that if the p-value is less than 0.005 for a certain gene, then the two Kaplan-Meier curves are statistically different, and therefore, the gene is a potential driver of prognosis.

3. Results

In this analysis, the main objective is to identify expression signatures that can separate primary and metastatic skin cutaneous melanoma (SKCM) based on RNA-seq expression data. After categorical features such as race, gender, ethnicity and vital status were converted by One Hot Encoding, the category Black and African American in the race data field was dropped since it has only one record and cannot be represented in both training and validation data set.

The data was random-stratified into training (70%) and validation sets (30%). Each of the models were fine-tuned using grid search scored using accuracy and F1 score.

One of the most widely used machine learning algorithms is Random Forest. Random Forest as feature selection falls under the category of embedded methods. In selecting biomarkers, each of the trees of random forest were built over a random extraction of patient observations from the TCGA data set and a random extraction of the genomic features. Since not every tree sees all of the characteristics or all of the findings, the trees are de-correlated and therefore less vulnerable to overfitting. Each tree estimator has a series of true or false questions based on the level of expression of each of the genes and divides the observations based on their respective similarities and differences. Thus, the ranking of importance of each gene came from how pure they are. The measure of impurity used in this study is the Gini impurity. To have a deeper understanding, features selected at the top of the trees are usually more important than features selected at the end nodes of the trees, since top splits generally result in larger knowledge gains.

Random Forest was trained first with 19815 genes and was fined-tuned using grid-search to find the optimal hyperparameters. The Random Forest model used for biomarker discovery has the following hyperparameters: maximum features of 60%, minimum samples of each leaf is 8 and the number of estimators is 30. Then, feature importance of the model was analyzed. Table 1 shows the top 30 genes out of 139 that has been found significant (i.e., weighted).

Table 1. The 30 genes exhibiting the highest scores through Random Forest for Biomarker Discovery analysis

Rank	Gene Code	Gene Name	Score
1	C7	Complement C7	0.1591
2	KRT17	Keratin 17	0.1029
3	CLEC2A	Keratinocyte-Associated C-Type Lectin	0.0912
4	S100A7A	S100 Calcium Binding Protein A7A	0.0646
5	KRTDAP	Keratinocyte Differentiation Associated Protein	0.0604
6	WFDC5	WAP Four-Disulfide Core Domain 5	0.0418
7	KRT6B	Keratin 6B	0.0389
8	S100A7	S100 Calcium-Binding Protein A7 (Psoriasin 1)	0.0242
9	KRT14	Keratin 14	0.0196
10	PVRL4	Nectin Cell Adhesion Molecule 4	0.0176
11	SERPINB4	Squamous Cell Carcinoma Antigen 2	0.0172
12	IL20RB	Interleukin 20 Receptor Subunit Beta	0.0114
13	AFAP1-AS1	AFAP1 Antisense RNA 1	0.0109
14	FKBP1B	FKBP Prolyl Isomerase 1B	0.0103
15	ZSWIM7	Zinc Finger SWIM-Type Containing 7	0.0094
16	PRG2	Proteoglycan 2, Pro Eosinophil Major Basic Protein	0.0091
17	PAX1	Paired Box Protein Pax-1	0.0087
18	DMBT1	Deleted In Malignant Brain Tumors 1	0.0086
19	ZNF653	Zinc Finger Protein 65	0.0085
20	GRIK5	Glutamate Ionotropic Receptor Kainate Type Subunit 5	0.0081
21	MMP3	Matrix Metalloproteinase 3	0.0080
22	ZNF593	Zinc Finger Protein 593	0.0075
23	VDAC1	Outer Mitochondrial Membrane Protein Porin 1	0.0073
24	ADAMTSL3	ADAMTS Like 3	0.0072
25	RGS4	Regulator Of G Protein Signaling 4	0.0071
26	MRPL44	Mitochondrial Ribosomal Protein L44	0.0070
27	LYSMD2	LysM Domain Containing 2	0.0068
28	TDRKH	Tudor And KH Domain Containing	0.0059
29	CSPG4	Melanoma-Associated Chondroitin Sulfate Proteoglycan	0.0057
30	PLA2G2F	Phospholipase A2 Group IIF	0.0056

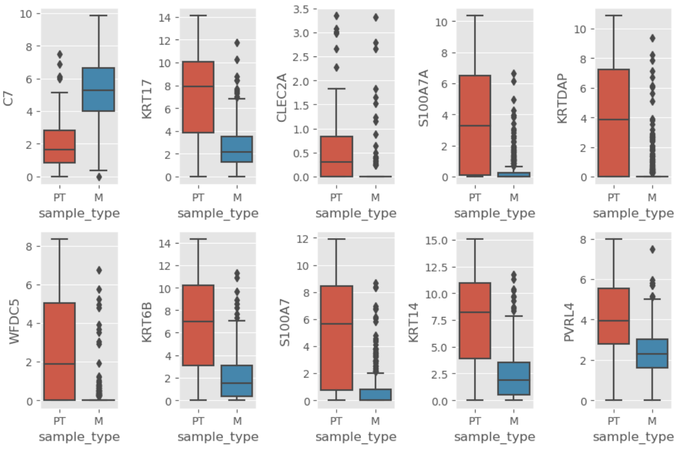


Figure 2. Box Plots of Top 10 Gene Expressions. M refers to Metastatic and PT refers to Primary Tumor

The expression of the Top 10 genes learnt from Random Forest as potential biomarkers was further examined as shown in Figure 2. C7 is upregulated in metastatic compared to primary tumor, while KRT17, CLEC2A, S1007A, KRTDAP, WFDC5, KRT6B, S100A7, KRT14 and PVRL4 are downregulated (or, upregulated in primary tumor compared to metastatic). The rest of the genes showing significant ($p < 0.05$, Welch's t-test, see Supplementary Materials File 10) upregulated expression in either primary tumor or metastatic sample type is shown in Table 2.

Table 2. Upregulated genes according to sample type from among the 139 feature selected genes

Primary Tumor	Metastatic
KRT17, CLEC2A, S100A7A, KRTDAP, WFDC5, KRT6B, S100A7, KRT14, PVRL4, SERPINB4, IL20RB, PAX1, MMP3, PLA2G2F, FCER1A, PSMD9, PRKRIP1, HMG20B, RAX, SSNA1, MRRF, PITHD1, COQ4, XKRX, FAM109B, C1orf159, MIEN1, RNF135, AKR1B15, SPSB3, SWI5, ATP12A, LCE1F, ALAD, FAAH, RDH12, RPS28, VDAC1, G6PC3, FAM98C, ZNF593, MRPL44, TBC1D13, ZSWIM7, PRG2, CICP27, CIB2, FKBP1B, ZNF653	C7, DOCK11, SCN4A, CLIC5, PDK4, SNAP23, PABPC4L, SMARCAL1, SAMD8, CCPG1, MRPL23, SLC9A8, TSPAN14, RARRES2, SLC40A1, GSR, IGF1R, DDX3X, PSTPIP2, CASK, SMTNL2, ADAMTSL3, ARHGAP22, RGS4, GTF2H2C, TAF5L, LYSMD2, TDRKH

Looking at the genomic features extracted, there are genes of related functionality such as KRT17, KRTDAP, KRT6B and KRT14. Random Forest showed the genes that have potential for classifying melanoma based on specific gene expression. We thought it important to look for interactions/connections in order to derive the genes that hold the most information. In developing a model, this can be viewed as optimizing the bias - variance trade-off, wherein high bias can miss possible relevant genes (underfitting) and high variance may include multicollinear genomic features (overfitting) in the model.

Table 3. Betweenness Centrality Rank of Genes from Protein-Protein Interaction Network

Rank	Gene Code	Gene Name	Score
1	PC	Pyruvate Carboxylase, Mitochondrial	0.0886
2	RPN2	Ribophorin II	0.0636
3	TSHR	Thyroid Stimulating Hormone Receptor	0.0490
4	GSR	Glutathione Reductase, Mitochondrial	0.0396
5	RPS28	Ribosomal Protein S28	0.0370
6	GRIK5	Glutamate Ionotropic Receptor Kainate Type Subunit 5	0.0185
7	GNG2	G Protein Subunit Gamma 2	0.0131
8	C7	Complement C7	0.0130
9	S100A7	S100 Calcium-Binding Protein A7 (Psoriasin 1)	0.0104
10	SERPINB4	Squamous Cell Carcinoma Antigen 2	0.0078
11	IGF1R	Insulin Like Growth Factor 1 Receptor	0.0062
12	KRT14	Keratin 14	0.0061
13	NKX6-1	NK6 Homeobox 1	0.0052
14	MRRF	Ribosome-Recycling Factor, Mitochondrial	0.0051
15	RPE65	Retinoid Isomerohydrolase RPE65	0.0045
16	LMX1B	LIM Homeobox Transcription Factor 1 Beta	0.0032
17	PAX1	Paired Box Protein Pax-1	0.0032
18	PTF1A	Pancreas Associated Transcription Factor 1a	0.0032
19	PTS	6-Pyruvoyltetrahydropterin Synthase	0.0026
20	KRT6B	Keratin 6B	0.0016
21	CASK	Calcium/Calmodulin Dependent Serine Protein Kinase	0.0013
22	FBXW10	F-Box And WD Repeat Domain Containing 10	0.0006

Out of the 139 genomic features identified by random forest, 22 genes were found to be highly connected with other genes on the list. Among these 22 information heavy genes, **C7, S100A7, SERPINB4, GRIK5, KRT14, PAX1, KRT6B** figured prominently in feature selection (Table 1), while genes such as **PC, RPN2, TSHR, GSR, RPS28, GNG2** which were not as prominent have risen to the top.

3.1. Model Performance

During the model tuning, F1 and accuracy scores were used as metrics to improve the performance since there is imbalanced class in the data set. AUC score was used as the final metric as it is commonly used to depict the trade-off relationship between clinical sensitivity and specificity for each potential cut-off for a test or a set of tests in a graphical format. Furthermore, AUC provides insight into the value of using the model in diagnosing melanoma patients. To wit, it determines how well the model is in correctly classifying a metastatic melanoma patient given the yield probability that the patient indeed has metastatic melanoma.

Table 4. Model Performance (Validation Dataset) using Biomarkers Discovered through Random Forest

	Model	Genes	F1	Accuracy	AUC
RF-RF	Random Forest	Top 20	92.85	93.01	0.9789
RF-SVM-R	SVM (Radial Basis Kernel)	Top 10	86.60	87.76	0.9249
RF-LR	Logistic Regression	Top 10	87.59	88.81	0.9234
RF-NB	Naive Bayes	Top 20	80.04	82.52	0.8252
RF-SVM-L	SVM (Linear Kernel)	Top 10	80.80	83.91	0.8205
RF-SVM-Sig	SVM (Sigmoid Kernel	Top 30	79.02	80.06	0.8054

The model performance were compared for the top genes, progressively selected (Top 10, Top 20, Top 30, Top 40, Top 50 as in [1]), using logistic regression, support vector machines (linear, polynomial, radial basis and sigmoid kernel), gaussian naive bayes, and random forest models. These models were trained using 5-fold cross validation. Overfitting occurs as performance on the training set improves but performance on the validation or test data set worsens; thus, the criteria to determine which best suitable number of genes to a specific algorithm is see the performance gap between training set and validation set. Unfortunately, support vector machines with polynomial kernel does not perform well with the 139 genes identified [see Supplementary Materials, File 3 - 4]. Table 4 shows the 6 best models and their validation scores that achieved high AUC scores. Based on the results, only the Top 30 (out of 139) identified genes were found to be important in diagnosing melanoma.

Table 5. Model Performance (Validation Dataset) using Biomarkers Discovered through Random Forest and Mapped by Protein-Protein Interaction Network

Model	Name	Genes	F1	Accuracy	AUC
RF-PPI-SVM-L	SVM (Linear Kernel)	Top 20	83.15	85.36	0.9659
RF-PPI-NB	Naive Bayes	Top 20	86.43	87.80	0.9054
RF-PPI-SVM-Sig	SVM (Sigmoid Kernel)	Top 10	73.17	79.67	0.9049
RF-PPI-LR	Logistic Regression	Top 10	83.15	85.37	0.8808

Likewise in the PPI mapped genes that were ranked using betweenness centrality, the genes were progressively selected (Top 10, Top 20) and the performance were compared. Table 5 shows that logistic regression, support vector machines and naive bayes achieved high validation AUC scores. [See Supplemetary Materials, File 5 - 6]

To further investigate the performance of the models in terms of their AUC scores, De Long's test [31] was conducted. It found that **RF-LR Top 10, RF-RF Top 20, RF-PPI-SVM-Sig Top 10** models were significantly better ($p < 0.05$). [See Supplemetary

Materials, File 7] The rank of Random Forest selected features might still miss some relevant genes and the rank using betweenness centrality might also increase the variance estimates across the samples. This can be further supported by the analysis on bias - variance decomposition among these models, found in Table S2, showing that **RF-LR Top 10, RF-RF Top 20, RF-PPI-SVM-Sig Top 10** models' expected loss were minimized as bias and variance were being optimized. Finally, the three significant models were combined as an ensemble model through soft-voting. The resulting ensemble model still has high and acceptable validation scores (F1 = 90.60, Accuracy = 91.05, AUC Score = 0.9861, see Supplementary Materials, Table S2), after making sure that the bias and variance were minimized. The unique gene signatures that were used in the ensemble model are listed in Table 6.

Table 6. The 26 genes in the Ensemble Meta Classifier with Soft Voting

Gene Code	Gene Name	Location
S100A7	S100 Calcium-Binding Protein A7 (Psoriasin 1)	chr1
S100A7A	S100 Calcium Binding Protein A7A	chr1
PVRL4	Nectin Cell Adhesion Molecule 4	chr1
FKBP1B	FKBP Prolyl Isomerase 1B	chr2
IL20RB	Interleukin 20 Receptor Subunit Beta	chr3
AFAP1-AS1	AFAP1 Antisense RNA 1	chr4
C7	Complement C7	chr5
GSR	Glutathione Reductase, Mitochondrial	chr8
DMBT1	Deleted In Malignant Brain Tumors 1	chr10
PRG2	Proteoglycan 2, Pro Eosinophil Major Basic Protein	chr11
PC	Pyruvate Carboxylase, Mitochondrial	chr11
CLEC2A	Keratinocyte-Associated C-Type Lectin	chr12
KRT6B	Keratin 6B	chr12
GNG2	G Protein Subunit Gamma 2	chr14
TSHR	Thyroid Stimulating Hormone Receptor	chr14
KRT14	Keratin 14	chr17
KRT17	Keratin 17	chr17
ZSWIM7	Zinc Finger SWIM-Type Containing 7	chr17
GRIK5	Glutamate Ionotropic Receptor Kainate Type Subunit 5	chr19
KRTDAP	Keratinocyte Differentiation Associated Protein	chr19
RPS28	Ribosomal Protein S28	chr19
ZNF653	Zinc Finger Protein 653	chr19
SERPINB4	Squamous Cell Carcinoma Antigen 2	chr18
PAX1	Paired Box Protein Pax-1	chr20
RPN2	Ribophorin II	chr20
WFDC5	WAP Four-Disulfide Core Domain 5	chr20

3.2. Kaplan-Meier Survival Analysis

We performed the Kaplan-Meier survival analysis on the 139 significant genes (Table 2) selected by Random Forest. Logrank test identified that out of the 139 genes, 26 display a significant difference (p-value less than 0.005) in terms of survival between high and low expressor individuals.

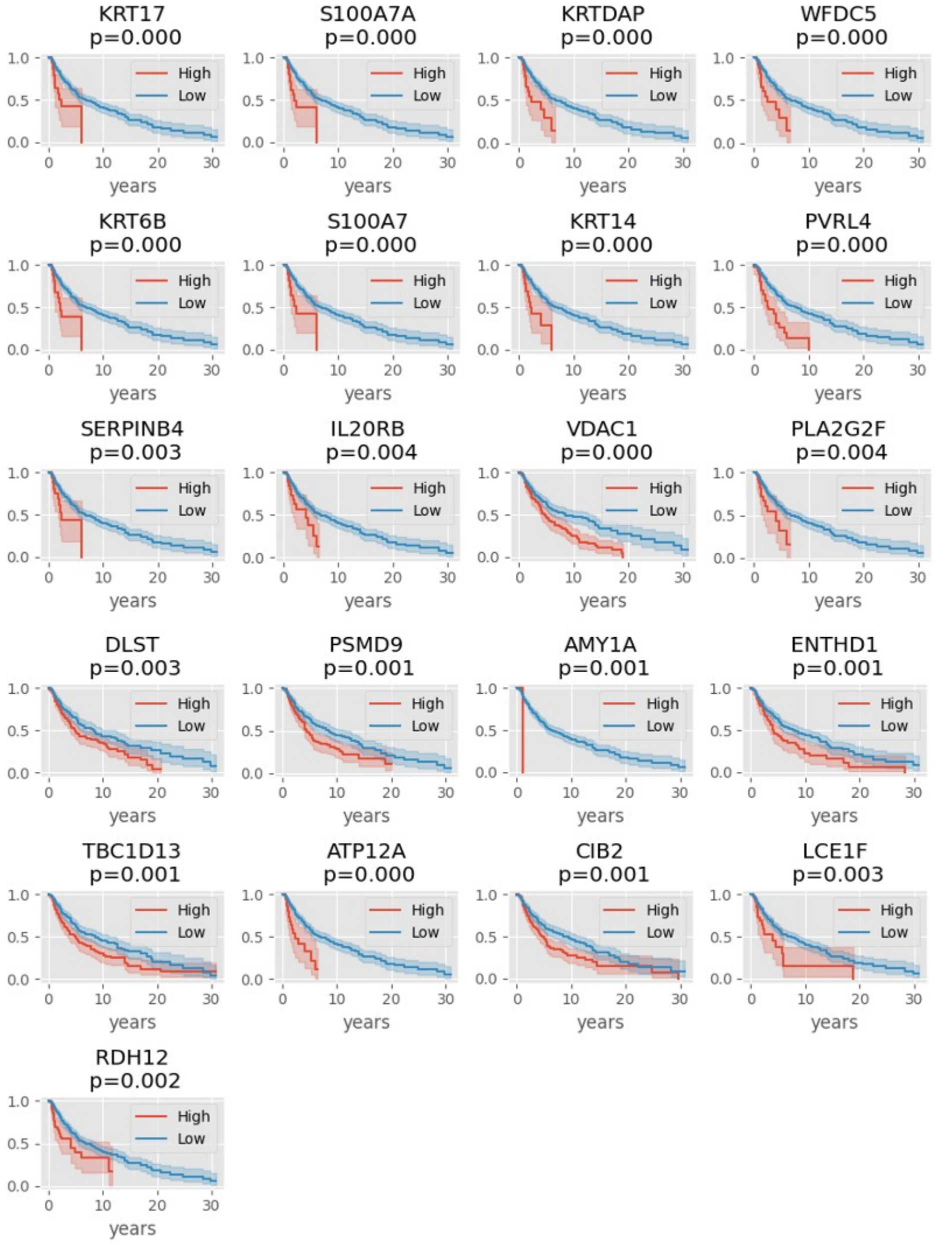


Figure 3. Kaplan-Meier plots showing that high expressors of these 21 genes have worse survival, where y-axis represents the probability of survival. In this analysis, the patient samples were divided in terms of the average of gene expression (high= above the mean, low = below the mean)

Our analysis shows that high expression of **KRT17**, **S100A7A**, **KRTDAP**, **WFDC5**, **KRT6B**, **S100A7**, **KRT14**, **PVRL4**, **SERPINB4**, **IL20RB**, VDAC1, PLA2G2F, DLST, PSMD9, AMY1A, ENTHD1, TBC1D13, ATP12A, CIB2, LCE1F, or RDH12 is associated with worse survival as shown in Figure 3.

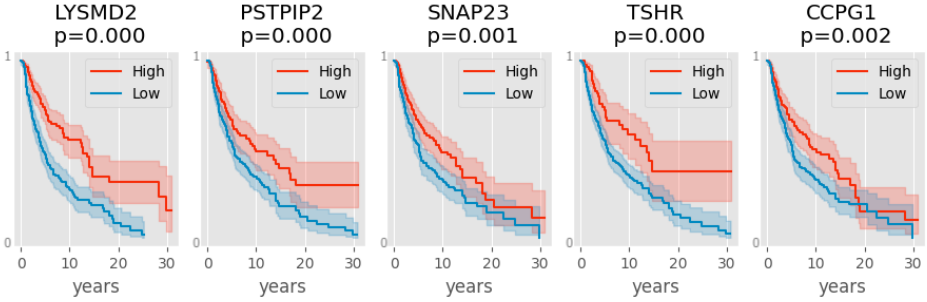


Figure 4. Kaplan-Meier plots showing that high expressors of these 5 genes have better survival, where y-axis represents the probability of survival. In this analysis, the patient samples were divided in terms of the average of gene expression (high= above the mean, low = below the mean)

On the other hand, high expression of LYSMD2, PSTPIP2, SNAP23, **TSHR**, or CCPG1 is associated with better survival as shown in Figure 4.

Eleven (11) of these 26 prognostic genes, in bold text, are common with the 26 genes identified by the ensemble classifier (Table 7).

3.3. Implicit Bias

The evaluation of data is a critical step in the development of machine learning models, especially when they are used in clinical decision support for medical diagnosis. Initial exploratory analysis show that training the model with the TCGA data for melanoma has implicit bias on race, gender, age groups, and BMI. There were more metastatic samples analyzed compared to primary tumor; patients were concentrated in the 40 - 79 age groups; the female to male ratio is 0.58 (Figure S2); in terms of race, whites were dominantly present in the data set (Figure S3); and for BMI, underweight patients were not represented (Figure S4). We hypothesized that our machine learning model will only perform well on populations that are well represented.

Table 7. Performance accuracy of the models to examine implicit bias in training dataset

Variable	N	Ensemble
Sample Type		
Primary Tumor	68	51.47%
Metastatic	218	99.54%
Age		
0-19	3	100.00%
20-39	29	89.66%
40-59	109	90.83%
60-79	124	87.09%
80+	21	71.19%
Gender		
Male	181	88.95%
Female	105	86.66%
BMI		
Normal	92	88.04%
Overweight	107	90.65%
Obese	86	84.88%
Race		
White	271	88.19%
Asian	10	80.00%

We calculated the performance accuracy on sample type, age groups, gender, BMI, and race using the ensemble model. Results shown in Table 7 confirmed our hypothesis for sample type and age groups, but not for gender and race where the model still performs quite well despite the unevenness of the data. For the age range 0-19, the perfect performance of the model is likely an overfit due to the very small number of samples.

4. Discussion

Feature selection in machine learning applied to gene expression data is a powerful method that can identify biomarkers to classify disease states (primary vs. metastatic melanoma in this study). Once the list of potential biomarkers are narrowed down and ranked for their respective contributions (139 weighted genes ranked), additional machine learning methods such as LR, SVM, and NB can further indicate which rank cut off is important based on model performance (in this case, Top 30) giving a more manageable set of molecular markers for further study.

Protein-protein interaction analysis of the original 139 genes in the RF learnt set, showed that lower ranked genes can also figure prominently, indicating that interactions may be important. Of the Top 10 genes in betweenness centrality score, only C7 and S100A7 were also in the RF Top 10 (see Table 1), yet the performance of models incorporating the Top 10 PPI-selected genes was still very high (0.935-0.9552, Table 5) even though PC, RPN2, TSHR, GSR, RPS28, and GNG2 genes were not even in the RF Top 30.

Comparison of the performance of several models distinguished the 3 best, which when ensembled, so as not to miss other relevant genes, performed very well (AUC=0.9861, see Supplementary Materials, Table SX). There are 26 genes in the ensemble meta-classifier, including genes involved in skin cell differentiation (CLEC2A, KRT6B, KRT14, KRT17, KRTDAP), immunity (S100A7, S100A7A, IL20RB, C7, PRG2, SERPINB4, WFDC5, FKBP1B), cell adhesion (PVRL4), energy/metabolism (PC, TSHR), cancer metastasis (AFAP1-AS1) and suppression (DMBT1), cellular redox (GSR), cell signalling (GNG2), cell division (ZSWIM7), protein synthesis and modification (RPS28, RPN2) and transcriptional regulation (ZNF653, PAX1). Most of these genes have also been linked to other cancers, thus, the methods we employed found genes involved in metastatic progression which could be common among cancers. Interestingly, GRIK5 (glutamate ionotropic receptor kainate type subunit 5), identified here for the first time as a classifier for primary vs. metastatic melanoma, is mainly known for its role in neural development and neuropsychiatric disorders [37–39].

Examining the profiles of the 139 genes in patient tissue, we found that some of these genes are highly expressed in metastatic tissue compared with primary tumor, such as C7, DOCK11, SCN4A etc. However, more genes in this set were expressed highly in primary tumors (Table 2). Genes such as members of the keratin family (KRT17, KRTDAP, KRT6B, KRT14, KRTAP13-2) are expressed more in primary tumor, perhaps indicating the differentiated status of less advanced cancers, or this could be a disruption in their normal expression by melanoma processes. (Unfortunately, we could not compare expression with normal skin tissue because there were none of these samples included in the data set.)

When expression of these genes were correlated with patient survival, we found genes whose high expression correlate with worse (Figure 3) or better (Figure 4) survival. Some of the genes that were highly expressed in primary tumors (such as the keratin genes) turned out to be predictive, but, oddly, of poor outcome. We can only surmise that perhaps the early stages melanoma ramps up the expression of these genes but this disruption may be detrimental to the patient eventually. Only 5 genes were found to be predictive of good outcome when highly expressed in melanoma: LYSMD2, PSPIP2, SNAP23, TSHR and CCPG1. Of these, only TSHR was identified by the ensemble classifier. TSHR (thyroid stimulating hormone receptor) controls thyroid cell metabolism, and defective TSHR causes hyperthyroidism. The expression of this hormone receptor has been observed in melanoma [40]; its downregulation has been associated with thyroid cancer metastasis and is prognostic for poor survival [41], in agreement with our findings. Moreover, TSHR has been identified for therapeutic intervention or as a theranostic indicator for thyroid, ovarian and hepatic malignacies [42], demonstrating the utility of our methods in the identification of potential therapeutic targets for oncology. Very little is known about LYSMD2 and PSPIP2, but SNAP23 (synaptosome associated protein 23) is a vesicular transport protein that is highly expressed in lymph nodes and the spleen (**https://www.ncbi.nlm.nih.gov/gene/8773**) pointing to a possible involvement in immunity. CCPG1 (cell cycle progression 1) is involved in endoplasmic reticulum homeostasis [43] and may be a tumor suppressor gene [2].

Assessing the data for potential bias is an exercise recommended and to be continuously conducted during AI implementation, in order to correct for under or over representation of specific populations in machine learning, and to help interpret model performance. The ultimate goal is to be able to roll out a fairer algorithm, which, if used in health, would not result in further inequities as is machine learning's wont. The analysis predetermines with what segments of the patient population our models would likely work in. Per our assessments, the TCGA SKCM data set is biased on sample type (metastatic > primary tumor), age (40-79 year olds are best represented), gender (male > female) and race (mostly white, few asians, and no other race categories). We expected our final model to perform best in the most represented groups which it did in terms of sample type and age. Surprisingly, the model still performed robustly with respect to gender and race. It must be tolerant down to a gender ratio of 0.58 (female to male), however, it is very interesting that even with an extreme race ratio (0.04, asian to white), the model still works albeit with somewhat lowered performance. We dropped the single black patient for this analysis, so we cannot generalize this model to the black population. The lack of data may be a reflection of the relatively low incidence of melanoma in blacks. For BMI, the segments are fairly represented, save that there were no underweight patients. Consequently, the model performed well in all BMI segments, but, again, it may not be extendable to underweight patients.

4.1. Biomarkers

Our models were able to identify notable genes, specifically ones also flagged by survival analysis. Some of these genes have been identified in previous studies involving machine learning on the same dataset [14,15]. These genes are good candidates for validation ahead of their potential applications in the clinical setting.

4.1.1. Potential Diagnostic Classifiers

C7. C7 is a member of the soluble Membrane Attack Complex (sMAC), together with C5b, C6, C8, and C9, generated upon activation of the complement system.[44] In a study done by Bhalla et al (2019) that used several feature selection methods on genomic data, C7 figured prominently in melanoma carcinogenesis and was also found to be upregulated in metastatic tumors. [15] Opposing observations were seen among ovarian and non-small cell lung cancer (NSCLC) tissues as C7 was found to be further downregulated as the tumor stage increased. More importantly, low C7 levels was also identified to be a significant prognostic factor for NSCLC patients. [45] The inclusion of C7 as a diagnostic classifier to distinguish between primary and metastatic melanoma is promising and warrants further investigation.

GRIK5. GRIK5 encodes for the kainate-preferring glutamate receptor subunit KA2 that is ubiquitously expressed in the mammalian brain. [46] In the SKCM dataset, the expression of GRIK5 does not seem to be significantly different between primary and metastatic melanoma. However, in preliminary studies on zebrafish, decreased expression of GRIK5 was found to lead to vascular pathologies in the eye and brain. They have been shown to be associated with patterning and vasculature integrity. [47] Given the earlier observations, the potential role of GRIK5 in angiogenic processes necessary for metastasis merits further investigation.

4.1.2. Potential Diagnostic and Prognostic Biomarkers

S100A7/S100A7A. In this study, S100A7 was shown to be upregulated in primary melanoma. An analysis of publicly-available gene expression profiles showed that S100A7 was highly expressed in primary cutaneous melanoma, but was significantly decreased in normal skin tissue and metastatic melanoma. A follow-up analysis of protein-protein interactions identified S100A7 as a hub gene in primary cutaneous melanoma. [48]

At the transcriptome level, a study done by Riker et al (2008), showed that S100A7 expression was highly expressed in primary cutaneous melanoma vs. normal skin tissue, but was seen to significantly decrease in metastatic melanoma. [10] A similar study showed several S100 family genes, including S100A7, being highly expressed in primary melanoma, but were seen to significantly decrease in metastatic melanoma. [49] More importantly, higher levels of S100A7 were detected in the urine of cutaneous melanoma patients compared to a control group. In addition, this trend was not seen in a heterogeneous group of patients with other cancer types. [50] The significant levels of S100A7 expression in primary cutaneous melanoma and the ease in detection in urine samples make it a promising diagnostic classifier.

KRT14, KRT17, KRT6B, KRTDAP. These genes are involved in keratinization. Increased expression of KRT6B, KRT14 and KRT17 were associated with poor suvival in melanoma. [51] KRTDAP, on the other hand, was found to have higher expression in primary tumour compared to metastatic tumour. [10] The role of KRTDAP is mainly in keratinocyte differentiation as such, this may indicate that metastatic melanoma tissue is less differentiated compared to primary lesions.

SERPINB4. Together with SERPINB3,SERPINB4 has been shown to be involved in inflammatory conditions of the skin and respiratory diseases such as Chronic Obstructive Pulmonary Disease (COPD) and tuberculosis [52]. It encodes for the Squamous Cell Carcinoma Antigen (SCCA) 2 which has been used as a diagnostic marker for advanced squamous cell carcinoma in the head and neck [53]. It has also been found to induce epithelial-mesenchymal transition (EMT) in mammalian epithelial cells, insinuating its role in tumor metastasis [54].

TSHR. TSHR has been documented to be expressed in all melanocytic lesions, with higher levels found in malignant and pre-malignant lesions. Its ligand, the thyroid stimulating hormone, was found to induce melanoma proliferation. Circulating levels of TSH increase in thyroid failure conditions providing an environment where melanoma can proliferate [55]. In the clinical setting, it has been documented that patients with cutaneous malignant melanoma are at a higher risk for other cancers, especially thyroid carcinoma [56].

PVRL4. PVRL4, also known as NECTIN4, has been identified as a potent inducer of anchorage-independent growth in epithelial cell culture [57]. In cancer, an increased expression of PVRL4 was found to be associated with high-grade serous ovarian carcinoma, but did not seem to be involved in survival [58].

WFDC5. WFDC5 is highly expressed in human epidermis [59] that are known to secrete protease inhibitors involved in inflammatory processes [60]. It was found to be upregulated in head and neck squamous cell carcinoma (HNSCC) expression data from the GEO database [61]. Using microarray data, WFDC5 figured in the top 40 of a candidate 200-gene signature that is able to distinguish between melanoma and normal epithelial cells / benign nevus. [62] .

IL20RB. IL20RA and IL20RB are subunits of the interleukin 20 receptor type I (IL20RI) found in the epidermis [63,64]. IL20RB was found to be associated with inflammatory processes in psoriasis [65] and vitiligo [66]. IL20RB expression levels have already been documented in several cancers. Cui et al (2019) showed that it is highly expressed in papillary renal cell carcinoma (PRCC) tissue, and was linked to poor prognosis among patients. In the same study, its repression limited the proliferation and migration of PRCC cells, thus highlighting its potential role in the EMT mechanisms leading to metastasis [67]. This finding can be corroborated by the function of one of the IL20R1 (IL20RA + IL20RB) / IL20R2 heterodimer ligand, IL20, which is a pro-inflammatory cytokine found to enhance wound healing, migration, and invasion in bladder cancer cell lines [68]. These evidence point to a potential role of IL20RB in inflammatory processes in melanoma which merits further investigation.

Author Contributions: Conceptualization, Jose Marie Antonio Miñoza, Manny Bacolod and Romulo de Castro; Data curation, Jose Marie Antonio Miñoza and Jonathan Adam Rico; Formal analysis, Jose Marie Antonio Miñoza, Jonathan Adam Rico, Pia Regina Fatima Zamora, Manny Bacolod, Reinhard Laubenbacher and Romulo de Castro; Funding acquisition, Romulo de Castro; Investigation, Pia Regina Fatima Zamora, Manny Bacolod and Romulo de Castro; Methodology, Jose Marie Antonio Miñoza and Jonathan Adam Rico; Project administration, Romulo de Castro; Resources, Jose Marie Antonio Miñoza, Jonathan Adam Rico, Manny Bacolod and Romulo de Castro; Software, Jose Marie Antonio Miñoza and Jonathan Adam Rico; Supervision, Manny Bacolod, Reinhard Laubenbacher and Romulo de Castro; Validation, Pia Regina Fatima Zamora, Manny Bacolod, Reinhard Laubenbacher and Romulo de Castro; Visualization, Jose Marie Antonio Miñoza and Jonathan Adam Rico; Writing – original draft, Jose Marie Antonio Miñoza, Jonathan Adam Rico, Pia Regina Fatima Zamora and Romulo de Castro; Writing – review and editing, Jose Marie Antonio Miñoza, Jonathan Adam Rico, Pia Regina Fatima Zamora, Manny Bacolod, Reinhard Laubenbacher and Romulo de Castro.

Funding: This research received support from the University of San Agustin and 3R Biosystems.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent with the use of TCGA data is covered by their Human Subjects Protection and Data Access Policies (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/policies/tcga-human-subjects-data-policies.pdf)

Data Availability Statement:

The data used in this study was downloaded from TCGA. (https://portal.gdc.cancer.gov/). The codes for this analysis can be found here: https://doi.org/10.5281/zenodo.4781962

Acknowledgments: The authors would like to thank Dr. Gerard Dumancas, Dominic Bolinas, Raphael Nelo Aguila and Dr. Vena Pearl Bongolan for valuable suggestions, support, and discussion. The results <published or shown> here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

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

Abbreviations

The following abbreviations are used in this manuscript:

TCGA	The Cancer Genome Atlas
SKCM	Skin Cutaneous Melanoma
M	Metastatic
PT	Primary Tumor
PPI	Protein - Protein Interaction
RF	Random Forest
LR	Logistic Regression
SVM	Support Vector Machines
NB	Naive Bayes

Supplementary Materials: Figure S1: Distribution of Sample Type. Figure S2: Distribution of Gender. Figure S3: Distribution of Race. Figure S4: Distribution of BMI. Figure S5-S34: Biomarker Discovery - Random Forest Estimators. Figure S35-S40: Box Plots of Top 139 Gene Expressions.

Table S1: Bias Variance Decomposition of the models. Table S2: Model Performance (Validation Dataset) of Ensemble Model with Soft Voting, Files: 01 Feature Selection (Biomarker Discovery) Results from Random Forest, 02 Feature Selection (Biomarker Discovery) Results from Random Forest mapped to PPI using Betweenness Centrality, 03 Model Evaluation Scores for Training Dataset [Based on Top 139 Genes of RF], 04 Model Evaluation Scores for Validation Dataset [Based on Top 139 Genes of RF], 05 Model Evaluation Scores for Training Dataset [Based on Top 20 Genes of RF-PPI], 06 Model Evaluation Scores for Validation Dataset [Based on Top 20 Genes of RF-PPI], 07 Model Comparison Results using De Long's Test for AUC, 08 Implicit Bias on Training Dataset [SKCM], 09 Bias Variance Decomposition, 10 Welchs T Test Results for 139 Genes

References

1. Wei, D. A multigene support vector machine predictor for metastasis of cutaneous melanoma. *Molecular Medicine Reports* **2017**. doi:10.3892/mmr.2017.8219.
2. Yang, S.; Xu, J.; Zeng, X. A six-long non-coding RNA signature predicts prognosis in melanoma patients. *International Journal of Oncology* **2018**. doi:10.3892/ijo.2018.4268.
3. Bennett, D.C. Review Article: How to make a melanoma: what do we know of the primary clonal events? *Pigment Cell & Melanoma Research* **2007**, *21*, 27–38. doi:10.1111/j.1755-148x.2007.00433.x.
4. Gray-Schopfer, V.; Wellbrock, C.; Marais, R. Melanoma biology and new targeted therapy. *Nature* **2007**, *445*, 851–857. doi:10.1038/nature05661.
5. Miller, A.J.; Mihm, M.C. Melanoma. *New England Journal of Medicine* **2006**, *355*, 51–65. doi: 10.1056/nejmra052166.
6. Braeuer, R.R.; Watson, I.R.; Wu, C.J.; Mobley, A.K.; Kamiya, T.; Shoshan, E.; Bar-Eli, M. Why is melanoma so metastatic? *Pigment Cell & Melanoma Research* **2013**, *27*, 19–36. doi: 10.1111/pcmr.12172.
7. Berk-Krauss, J.; Stein, J.A.; Weber, J.; Polsky, D.; Geller, A.C. New Systematic Therapies and Trends in Cutaneous Melanoma Deaths Among US Whites, 1986–2016. *American Journal of Public Health* **2020**, *110*, 731–733. doi:10.2105/ajph.2020.305567.
8. Mason, R.; Au, L.; Garces, A.I.; Larkin, J. Current and emerging systemic therapies for cutaneous metastatic melanoma. *Expert Opinion on Pharmacotherapy* **2019**, *20*, 1135–1152. doi: 10.1080/14656566.2019.1601700.
9. Kabbarah, O.; Nogueira, C.; Feng, B.; Nazarian, R.M.; Bosenberg, M.; Wu, M.; Scott, K.L.; Kwong, L.N.; Xiao, Y.; Cordon-Cardo, C.; Granter, S.R.; Ramaswamy, S.; Golub, T.; Duncan, L.M.; Wagner, S.N.; Brennan, C.; Chin, L. Integrative Genome Comparison of Primary and Metastatic Melanomas. *PLoS ONE* **2010**, *5*, e10770. doi:10.1371/journal.pone.0010770.
10. Riker, A.I.; Enkemann, S.A.; Fodstad, O.; Liu, S.; Ren, S.; Morris, C.; Xi, Y.; Howell, P.; Metge, B.; Samant, R.S.; Shevde, L.A.; Li, W.; Eschrich, S.; Daud, A.; Ju, J.; Matta, J. The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. *BMC Medical Genomics* **2008**, *1*. doi:10.1186/1755-8794-1-13.
11. Acs, B.; Rantalainen, M.; Hartman, J. Artificial intelligence as the next step towards precision pathology. *Journal of Internal Medicine* **2020**, *288*, 62–81. doi:10.1111/joim.13030.
12. Haenssle, H.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; Uhlmann, L.; Alt, C.; Arenbergerova, M.; Bakos, R.; Baltzer, A.; Bertlich, I.; Blum, A.; Bokor-Billmann, T.; Bowling, J.; Braghiroli, N.; Braun, R.; Buder-Bakhaya, K.; Buhl, T.; Cabo, H.; Cabrijan, L.; Cevic, N.; Classen, A.; Deltgen, D.; Fink, C.; Georgieva, I.; Hakim-Meibodi, L.E.; Hanner, S.; Hartmann, F.; Hartmann, J.; Haus, G.; Hoxha, E.; Karls, R.; Koga, H.; Kreusch, J.; Lallas, A.; Majenka, P.; Marghoob, A.; Massone, C.; Mekokishvili, L.; Mestel, D.; Meyer, V.; Neuberger, A.; Nielsen, K.; Oliviero, M.; Pampena, R.; Paoli, J.; Pawlik, E.; Rao, B.; Rendon, A.; Russo, T.; Sadek, A.; Samhaber, K.; Schneiderbauer, R.; Schweizer, A.; Toberer, F.; Trennheuser, L.; Vlahova, L.; Wald, A.; Winkler, J.; Wölbing, P.; Zalaudek, I. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology* **2018**, *29*, 1836–1842. doi:10.1093/annonc/mdy166.
13. Safran, T.; Viezel-Mathieu, A.; Corban, J.; Kanevsky, A.; Thibaudeau, S.; Kanevsky, J. Machine learning and melanoma: The future of screening. *Journal of the American Academy of Dermatology* **2018**, *78*, 620–621. doi:10.1016/j.jaad.2017.09.055.
14. Li, Y.; Krahn, J.M.; Flake, G.P.; Umbach, D.M.; Li, L. Toward predicting metastatic progression of melanoma based on gene expression data. *Pigment Cell & Melanoma Research* **2015**, *28*, 453–463. doi:10.1111/pcmr.12374.

15. Bhalla, S.; Kaur, H.; Dhall, A.; Raghava, G.P.S. Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients. *Scientific Reports* **2019**, *9*. doi:10.1038/s41598-019-52134-4.
16. Torrey, L.; Shavlik, J. Transfer Learning. In *Handbook of Research on Machine Learning Applications and Trends*; IGI Global, 2010; pp. 242–264. doi:10.4018/978-1-60566-766-9.ch011.
17. Ganchev, P.; Malehorn, D.; Bigbee, W.L.; Gopalakrishnan, V. Transfer learning of classification rules for biomarker discovery and verification from molecular profiling studies. *Journal of Biomedical Informatics* **2011**, *44*, S17–S23. doi:10.1016/j.jbi.2011.04.009.
18. Reif, D.M.; Motsinger, A.A.; McKinney, B.A.; Crowe, J.E.; Moore, J.H. Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types. 2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology. IEEE, 2006. doi:10.1109/cibcb.2006.330987.
19. Rivas, J.D.L.; Fontanillo, C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology* **2010**, *6*, e1000807. doi:10.1371/journal.pcbi.1000807.
20. Sevimoglu, T.; Arga, K.Y. The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal* **2014**, *11*, 22–27. doi:10.1016/j.csbj.2014.08.008.
21. Jaeger, S.; Aloy, P. From protein interaction networks to novel therapeutic strategies. *IUBMB Life* **2012**, *64*, 529–537. doi:10.1002/iub.1040.
22. Fionda, V. Networks in Biology. In *Encyclopedia of Bioinformatics and Computational Biology*; Elsevier, 2019; pp. 915–921. doi:10.1016/b978-0-12-809633-8.20420-2.
23. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; Jensen, L.J.; von Mering, C. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **2018**, *47*, D607–D613. doi:10.1093/nar/gky1131.
24. Mounir, M.; Lucchetta, M.; Silva, T.C.; Olsen, C.; Bontempi, G.; Chen, X.; Noushmehr, H.; Colaprico, A.; Papaleo, E. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLOS Computational Biology* **2019**, *15*, e1006701. doi:10.1371/journal.pcbi.1006701.
25. Silva, T.C.; Colaprico, A.; Olsen, C.; DAngelo, F.; Bontempi, G.; Ceccarelli, M.; Noushmehr, H. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research* **2016**, *5*, 1542. doi:10.12688/f1000research.8923.2.
26. Ayana, G.; Dese, K.; woon Choe, S. Transfer Learning in Breast Cancer Diagnoses via Ultrasound Imaging. *Cancers* **2021**, *13*, 738. doi:10.3390/cancers13040738.
27. Pardamean, B.; Cenggoro, T.W.; Rahutomo, R.; Budiarto, A.; Karuppiah, E.K. Transfer Learning from Chest X-Ray Pre-trained Convolutional Neural Network for Learning Mammogram Data. *Procedia Computer Science* **2018**, *135*, 400–407. doi:10.1016/j.procs.2018.08.190.
28. Noorbakhsh, J.; Farahmand, S.; pour, A.F.; Namburi, S.; Caruana, D.; Rimm, D.; Soltanieh-ha, M.; Zarringhalam, K.; Chuang, J.H. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nature Communications* **2020**, *11*. doi:10.1038/s41467-020-20030-5.
29. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. doi:10.1038/nature21056.
30. Gurung, S.; Gao, Y.R. Classification of Melanoma (Skin Cancer) using Convolutional Neural Network. 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA). IEEE, 2020. doi:10.1109/citisia50690.2020.9371829.
31. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, *44*, 837–845.
32. Sun, X.; Xu, W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters* **2014**, *21*, 1389–1393. doi:10.1109/lsp.2014.2337313.
33. Raschka, S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software* **2018**, *3*, 638. doi:10.21105/joss.00638.

34. Stalpers, L.; Kaplan, E. Edward L. Kaplan and the Kaplan-Meier Survival Curve. *BSHM Bulletin: Journal of the British Society for the History of Mathematics* **2018**, *33*, 109–135. doi: 10.1080/17498430.2018.1450055.
35. Kaplan, E.; Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **1958**, *53*, 457–481. doi:10.1080/01621459.1958.10501452.
36. Davidson-Pilon, C.; Kalderstam, J.; Jacobson, N.; Reed, S.; et. al.. CamDavidson-Pilon/lifelines: v0.25.11. *Zenodo* **2021**. doi:10.5281/zenodo.4683730.
37. Shibata, H.; Aramaki, T.; Sakai, M.; Ninomiya, H.; Tashiro, N.; Iwata, N.; Ozaki, N.; Fukumaki, Y. Association study of polymorphisms in the GluR7, KA1 and KA2 kainate receptor genes (GRIK3, GRIK4, GRIK5) with schizophrenia. *Psychiatry Research* **2006**, *141*, 39–51. doi: 10.1016/j.psychres.2005.07.015.
38. Gratacòs, M.; Costas, J.; de Cid, R.; Bayés, M.; González, J.R.; Baca-García, E.; de Diego, Y.; Fernández-Aranda, F.; Fernández-Piqueras, J.; Guitart, M.; Martín-Santos, R.; Martorell, L.; Menchón, J.M.; Roca, M.; Sáiz-Ruiz, J.; Sanjuán, J.; Torrens, M.; Urretavizcaya, M.; Valero, J.; Vilella, E.; Estivill, X.; and, Á.C. Identification of new putative susceptibility genes for several psychiatric disorders by association analysis of regulatory and non-synonymous SNPs of 306 genes involved in neurotransmission and neurodevelopment. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **2009**, *150B*, 808–816. doi:10.1002/ajmg.b.30902.
39. Yosifova, A.; Mushiroda, T.; Kubo, M.; Takahashi, A.; Kamatani, Y.; Kamatani, N.; Stoianov, D.; Vazharova, R.; Karachanak, S.; Zaharieva, I.; Dimova, I.; Hadjidekova, S.; Milanova, V.; Madjirova, N.; Gerdjikov, I.; Tolev, T.; Poryazova, N.; ODonovan, M.C.; Owen, M.J.; Kirov, G.; Toncheva, D.; Nakamura, Y. Genome-wide association study on bipolar disorder in the Bulgarian population. *Genes, Brain and Behavior* **2011**, *10*, 789–797. doi:10.1111/j.1601-183x.2011.00721.x.
40. Ellerhorst, J.A.; Sendi-Naderi, A.; Johnson, M.K.; Cooke, C.P.; Dang, S.M.; Diwan, A.H. Human melanoma cells express functional receptors for thyroid-stimulating hormone. *Endocrine-Related Cancer* **2006**, *13*, 1269–1277. doi:10.1677/erc.1.01239.
41. Liu, T.; Men, Q.; Su, X.; Chen, W.; Zou, L.; Li, Q.; Song, M.; Ouyang, D.; Chen, Y.; Li, Z.; Fu, X.; Yang, A. Downregulated expression of TSHR is associated with distant metastasis in thyroid cancer. *Oncology Letters* **2017**. doi:10.3892/ol.2017.7122.
42. Chu, Y.D.; Yeh, C.T. The Molecular Function and Clinical Role of Thyroid Stimulating Hormone Receptor in Cancer Cells. *Cells* **2020**, *9*, 1730. doi:10.3390/cells9071730.
43. Smith, M.D.; Wilkinson, S. CCPG1, a cargo receptor required for reticulophagy and endoplasmic reticulum proteostasis. *Autophagy* **2018**, pp. 1–2. doi:10.1080/15548627.2018.1441473.
44. Barnum, S.R.; Bubeck, D.; Schein, T.N. Soluble Membrane Attack Complex: Biochemistry and Immunobiology. *Frontiers in Immunology* **2020**, *11*. doi:10.3389/fimmu.2020.585108.
45. Ying, L.; Zhang, F.; Pan, X.; Chen, K.; Zhang, N.; Jin, J.; Wu, J.; Feng, J.; Yu, H.; Jin, H.; Su, D. Complement component 7 (C7), a potential tumor suppressor, is correlated with tumor progression and prognosis. *Oncotarget* **2016**, *7*, 86536–86546.
46. Hayes, D.; Braud, S.; Hurtado, D.; McCallum, J.; Sandley, S.; Isaac, J.; Roche, K. Trafficking and surface expression of the glutamate receptor subunit, KA2. *Biochem Biophys Res Commun* **2003**, *310*, 8–13. doi:10.1016/j.bbrc.2003.08.115.
47. Unlu, G.; Gamazon, E.R.; Qi, X.; Levic, D.S.; Bastarache, L.; Denny, J.C.; Roden, D.M.; Mayzus, I.; Breyer, M.; Zhong, X.; Konkashbaev, A.I.; Rzhetsky, A.; Knapik, E.W.; Cox, N.J. GRIK5 Genetically Regulated Expression Associated with Eye and Vascular Phenomes: Discovery through Iteration among Biobanks, Electronic Health Records, and Zebrafish. *The American Journal of Human Genetics* **2019**, *104*, 503–519. doi:10.1016/j.ajhg.2019.01.017.
48. H, D.; L, G.; M, L.; Z, C.; J, L.; J, T.; X, H.; Y, H.; K, X. Comprehensive analysis and identification of key genes and signaling pathways in the occurrence and metastasis of cutaneous melanoma. *Peerj* **2020**, *8*. doi:10.7717/peerj.10265.
49. feng Xiong, T.; qiang Pan, F.; Li, D. Expression and clinical significance of S100 family genes in patients with melanoma. *Melanoma Research* **2019**, *29*, 23–29. doi: 10.1097/CMR.0000000000000512.
50. Brouard, M.; Saurat, J.; Ghanem, G.; Siegenthaler, G. Urinary excretion of epidermal-type fatty acid-binding protein and S100A7 protein in patients with cutaneous melanoma. *Melanoma Research* **2002**, *12*, 627–631. doi:10.1097/00008390-200212000-00013.
51. Han, W.; Hu, C.; Fan, Z.J.; Shen, G.L. Transcript levels of keratin 1/5/6/14/15/16/17 as potential prognostic indicators in melanoma patients. *Scientific Reports* **2021**, *11*. doi: 10.1038/s41598-020-80336-8.

52. Sun, Y.; Sheshadri, N.; Zong, W.X. SERPINB3 and B4: From biochemistry to biology. *Seminars in Cell & Developmental Biology* **2017**, *62*, 170–177. doi:10.1016/j.semcdb.2016.09.005.
53. SAIDAK, Z.; MORISSE, M.C.; CHATELAIN, D.; SAUZAY, C.; HOUESSINON, A.; GUILAIN, N.; SOYEZ, M.; CHAUFFERT, B.; DAKPÉ, S.; GALMICHE, A. Squamous Cell Carcinoma Antigen-encoding Genes SERPINB3/B4 as Potentially Useful Markers for the Stratification of HNSCC Tumours. *Anticancer Research* **2018**, *38*, 1343–1352, [https://ar.iiarjournals.org/content/38/3/1343.full.pdf].
54. Sheshadri, N.; Catanzaro, J.M.; Bott, A.J.; Sun, Y.; Ullman, E.; Chen, E.I.; Pan, J.A.; Wu, S.; Crawford, H.C.; Zhang, J.; Zong, W.X. SCCA1/SERPINB3 Promotes Oncogenesis and Epithelial–Mesenchymal Transition via the Unfolded Protein Response and IL6 Signaling. *Cancer Research* **2014**, *74*, 6318–6329. doi:10.1158/0008-5472.can-14-0798.
55. Ellerhorst, J.; Cooksley, C.; Broemeling, L.; Johnson, M.; Grimm, E. High prevalence of hypothyroidism among patients with cutaneous melanoma. *Oncology Reports* **2003**. doi:10.3892/or.10.5.1317.
56. Kim, C.Y.; Lee, S.H.; Oh, C.W. Cutaneous Malignant Melanoma Associated with Papillary Thyroid Cancer. *Annals of Dermatology* **2010**, *22*, 370. doi:10.5021/ad.2010.22.3.370.
57. Pavlova, N.N.; Pallasch, C.; Elia, A.E.; Braun, C.J.; Westbrook, T.F.; Hemann, M.; Elledge, S.J. A role for PVRL4-driven cell–cell interactions in tumorigenesis. *eLife* **2013**, *2*. doi:10.7554/elife.00358.
58. Bekos, C.; Muqaku, B.; Dekan, S.; Horvat, R.; Polterauer, S.; Gerner, C.; Aust, S.; Pils, D. NECTIN4 (PVRL4) as Putative Therapeutic Target for a Specific Subtype of High Grade Serous Ovarian Cancer—An Integrative Multi-Omics Approach. *Cancers* **2019**, *11*, 698. doi:10.3390/cancers11050698.
59. Kalinina, P.; Vorstandlechner, V.; Buchberger, M.; Eckhart, L.; Lengauer, B.; Golabi, B.; Laggner, M.; Hiess, M.; Sterniczky, B.; Födinger, D.; Petrova, E.; Elbe-Bürger, A.; Beer, L.; Hovnanian, A.; Tschachler, E.; Mildner, M. The Whey Acidic Protein WFDC12 Is Specifically Expressed in Terminally Differentiated Keratinocytes and Regulates Epidermal Serine Protease Activity. *Journal of Investigative Dermatology* **2021**, *141*, 1198–1206.e13. doi:10.1016/j.jid.2020.09.025.
60. Gerber, P.A.; Hevezi, P.; Buhren, B.A.; Martinez, C.; Schrumpf, H.; Gasis, M.; Grether-Beck, S.; Krutmann, J.; Homey, B.; Zlotnik, A. Systematic Identification and Characterization of Novel Human Skin-Associated Genes Encoding Membrane and Secreted Proteins. *PLoS ONE* **2013**, *8*, e63949. doi:10.1371/journal.pone.0063949.
61. Sun, Y.; Zhang, Q.; Yao, L.; Wang, S.; Zhang, Z. Comprehensive analysis reveals novel gene signature in head and neck squamous cell carcinoma: predicting is associated with poor prognosis in patients. *Translational Cancer Research* **2020**, *9*, 5882–5892. doi:10.21037/tcr-20-805.
62. Liu, W.; Peng, Y.; Tobin, D.J. A new 12-gene diagnostic biomarker signature of melanoma revealed by integrated microarray analysis. *PeerJ* **2013**, *1*, e49. doi:10.7717/peerj.49.
63. Blumberg, H.; Conklin, D.; Xu, W.; Grossmann, A.; Brender, T.; Carollo, S.; Eagan, M.; Foster, D.; Haldeman, B.A.; Hammond, A.; Haugen, H.; Jelinek, L.; Kelly, J.D.; Madden, K.; Maurer, M.F.; Parrish-Novak, J.; Prunkard, D.; Sexson, S.; Sprecher, C.; Waggie, K.; West, J.; Whitmore, T.E.; Yao, L.; Kuechle, M.K.; Dale, B.A.; Chandrasekher, Y.A. Interleukin 20. *Cell* **2001**, *104*, 9–19. doi:10.1016/s0092-8674(01)00187-8.
64. Parrish-Novak, J.; Xu, W.; Brender, T.; Yao, L.; Jones, C.; West, J.; Brandt, C.; Jelinek, L.; Madden, K.; McKernan, P.A.; Foster, D.C.; Jaspers, S.; Chandrasekher, Y.A. Interleukins 19, 20, and 24 Signal through Two Distinct Receptor Complexes. *Journal of Biological Chemistry* **2002**, *277*, 47517–47523. doi:10.1074/jbc.m205114200.
65. Kingo, K.; Mössner, R.; Rätsep, R.; Raud, K.; Krüger, U.; Silm, H.; Vasar, E.; Reich, K.; Kõks, S. Association analysis of IL20RA and IL20RB genes in psoriasis. *Genes & Immunity* **2008**, *9*, 445–451. doi:10.1038/gene.2008.36.
66. Reimann, E.; Kingo, K.; Karelson, M.; Reemann, P.; Loite, U.; Sulakatko, H.; Keermann, M.; Raud, K.; Abram, K.; Vasar, E.; Silm, H.; Kõks, S. The mRNA expression profile of cytokines connected to the regulation of melanocyte functioning in vitiligo skin biopsy samples and peripheral blood mononuclear cells. *Human Immunology* **2012**, *73*, 393–398. doi:10.1016/j.humimm.2012.01.011.
67. Cui, X.F.; Cui, X.G.; Leng, N. Overexpression of interleukin-20 receptor subunit beta (IL20RB) correlates with cell proliferation, invasion and migration enhancement and poor prognosis

in papillary renal cell carcinoma. *Journal of Toxicologic Pathology* **2019**, *32*, 245–251. doi: 10.1293/tox.2019-0017.

68. Lee, S.J.; Lee, E.J.; Kim, S.K.; Jeong, P.; Cho, Y.H.; Yun, S.J.; Kim, S.; Kim, G.Y.; Choi, Y.H.; Cha, E.J.; Kim, W.J.; Moon, S.K. Identification of Pro-Inflammatory Cytokines Associated with Muscle Invasive Bladder Cancer; The Roles of IL-5, IL-20, and IL-28A. *PLoS ONE* **2012**, *7*, e40267. doi:10.1371/journal.pone.0040267.