


OPEN

# Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients

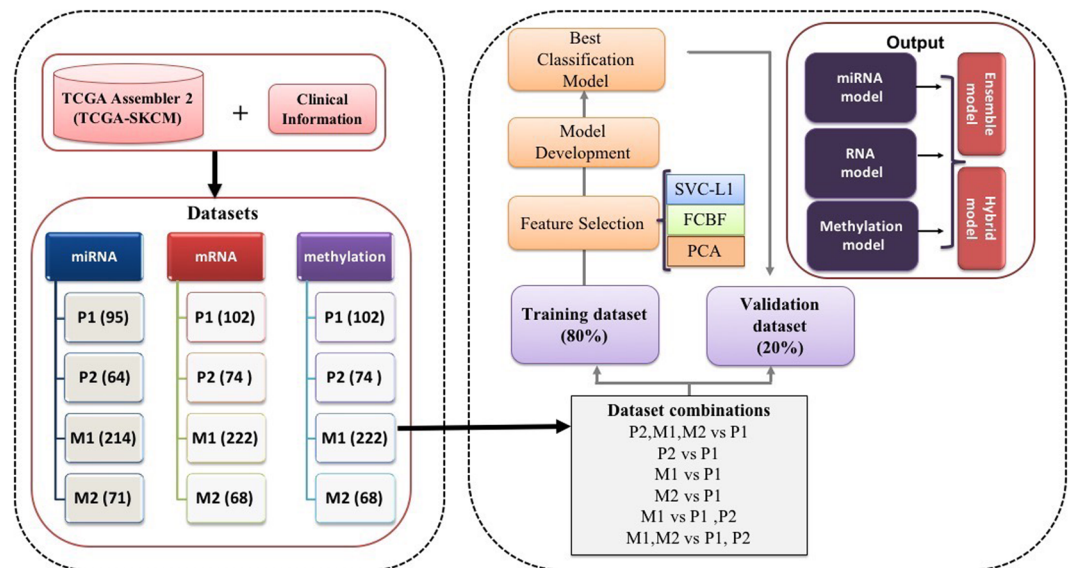
Sherry Bhalla<sup>1,2,4</sup>, Harpreet Kaur<sup>3,4</sup> , Anjali Dhali<sup>1</sup> & Gajendra P. S. Raghava<sup>1\*</sup>

The metastatic Skin Cutaneous Melanoma (SKCM) has been associated with diminished survival rates and high mortality rates worldwide. Thus, segregating metastatic melanoma from the primary tumors is crucial to employ an optimal therapeutic strategy for the prolonged survival of patients. The SKCM mRNA, miRNA and methylation data of TCGA is comprehensively analysed to recognize key genomic features that can segregate metastatic and primary tumors. Further, machine learning models have been developed using selected features to distinguish the same. The Support Vector Classification with Weight (SVC-W) model developed using the expression of 17 mRNAs achieved Area under the Receiver Operating Characteristic (AUROC) curve of 0.95 and an accuracy of 89.47% on an independent validation dataset. This study reveals the genes *C7*, *MMP3*, *KRT14*, *LOC642587*, *CASP7*, *S100A7* and miRNAs hsa-mir-205 and hsa-mir-203b as the key genomic features that may substantially contribute to the oncogenesis of melanoma. Our study also proposes genes *ESM1*, *NFATC3*, *C7orf4*, *CDK14*, *ZNF827*, and *ZSWIM7* as novel putative markers for cutaneous melanoma metastasis. The major prediction models and analysis modules to predict metastatic and primary tumor samples of SKCM are available from a webserver, CancerSPP (<http://webs.iitd.edu.in/raghava/cancerspp>).

Cancer is one of the major causes of mortality worldwide since the last few decades. According to GLOBOCAN, 2018, 18.1 million new cancer cases and 9.6 million deaths have been estimated worldwide. The melanoma contributes 1.6% of the new cancer cases and 0.6% of deaths due to cancer worldwide<sup>1</sup>. As per the American Cancer Society, there is an estimation of 96,480 melanoma related new cases and 7,230 deaths in 2019 in the US. Melanoma is more prominent in males as compared to females<sup>2</sup>. The malignant transformation of normal human epithelial melanocytes, located within the basement membrane of the skin results in melanoma development. There are several genetic<sup>3</sup> and environmental factors such as excessive exposure of UV radiations, indoor tanning devices and contacts with certain chemicals like arsenic and hydrocarbons, *etc.* that contribute to melanoma carcinogenesis<sup>4</sup>.

Recently, with the advancement of genomic technologies, there is a huge increment in the generation of big multi-omics data, particularly in the field of cancer<sup>5</sup>, which can be explored for the identification of diagnostic and prognostic cancer biomarkers<sup>6</sup>. The Cancer Genome Atlas (TCGA) is one of the prominent and inclusive repository containing genomic, transcriptomic, epigenetic, proteomics and clinical information of 33 types of cancer<sup>7</sup>. The core study on SKCM done by TCGA has revealed four subtypes of cancer, which include mutant *BRAF*, mutant *RAS*, mutant *NF1*, triple WT (wild-type) based on mutant genes. The triple WT SKCM subtype mainly exhibits *KIT* mutations, focal amplifications and structural rearrangements. Further, it has been observed that the mutational rate of these genes is much higher in melanoma patients than other cancer types of TCGA<sup>8,9</sup>. Interestingly, over 50% of melanoma patients have *BRAF* kinase (*BRAF* proto-oncogene, serine/threonine kinase) mutations<sup>10</sup>. In addition, various studies have demonstrated that the SKCM arises from the anomalies in transcriptomic and epigenetic factors such as expression of mRNAs, miRNAs, the aberration in methylation patterns of CpG islands of genes and histone modifications, which paves the way for the development of potential molecular biomarkers in melanoma<sup>11–23</sup>. In the past, several reports have revealed the potential role of miRNA expression as prognostic biomarkers in cutaneous melanoma. For instance, miR205 and miR29c both act as tumor suppressors and down-regulate the expression of *E2F1*, *E2F5*<sup>24</sup> and *DNMT3*<sup>15</sup> genes, respectively. Besides

<sup>1</sup>Center for Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India. <sup>2</sup>Centre for Systems Biology and Bioinformatics, Panjab University, Chandigarh, India. <sup>3</sup>CSIR-Institute of Microbial Technology, Chandigarh, India. <sup>4</sup>These authors contributed equally: Sherry Bhalla and Harpreet Kaur. \*email: [raghava@iitd.ac.in](mailto:raghava@iitd.ac.in)



**Figure 1.** The workflow of the study.

miRNAs, histone methyltransferases also act as crucial players in the progression of melanoma by enhancing the expression of enhancer of zeste homolog 2 (*EZH2*)<sup>25</sup>.

Earlier studies have scrutinized the distinctions between primary melanoma and metastatic melanoma<sup>26–32</sup>. The metastasis mechanism involves several pathways including epithelial-mesenchymal transition (EMT), angiogenesis and invasion. Furthermore, the aggressive stage of melanoma can metastasize to lymph nodes, distinct tissues and organs<sup>33</sup>. Although the survival of cutaneous melanoma patients is affected by various factors, the disease's early stage diagnosis is one of the most vital parameters with the greatest impact on survival. Different studies have shown the metastasis-free malignant melanoma patients, *i.e.* patients with primary tumor have significantly prolonged survival<sup>34,35</sup>. Evidently, the five-year relative survival rate of melanoma patients is 23%, 64% and 98% for distinct, regional stage and localized tumors, respectively<sup>2</sup>. Hence, the detection of tumor at a localized stage, *i.e.* primary tumor is crucial for patient management and implementation of an appropriate therapeutic strategy for prolonged survival of patients. The genomic and epigenomic biomarkers that can detect primary tumor with high precision might prove to be a boon in this regard and can eventually result in the better outcome of the patients with personalized treatment.

Previously, several stochastic stage wise prediction and classification methods have been developed for diverse cancer types<sup>36,37</sup>. Recently, one study has predicted the metastatic progression score for the assignment of metastatic and primary melanoma based on key miRNA and mRNA expression based putative biomarkers<sup>38</sup>. Although, all the metastatic samples were correctly assigned to a metastatic category based on metastatic progression score. But, the lack of gold standard performance measures like sensitivity, specificity and AUROC, absence of the performance on independent validation dataset and the unavailability of any web-service to analyse new data based on those identified markers are the major lacunae. Hence, the current study is designated to overcome these inadequacies.

In this analysis, we have made an effort to understand the cutaneous skin melanoma progression based on multi-omics layers of data in TCGA that comprises of RNAseq, miRNAseq and methylation expression. Through state-of-the-art machine learning-based feature selection techniques, we have identified genomic signatures that can categorize both primary and metastatic samples with high accuracy. Subsequently, prediction models were developed based on these key identified genomic features using several supervised machine learning techniques that can segregate primary and metastasized SKCM patients.

## Results

In the current study, we have analysed the RNAseq, miRNAseq, methylation-seq data of SKCM from TCGA for 466, 444, 466 patients, respectively. To mine important genomic and epigenomic features which can discriminate various degree of metastatic tumors (P2, M1 and M2) from the primary tumors (P1), we used well established feature selection methods like WEKA-FCBF<sup>39,40</sup>, Support vector machines with L1 regularization (SVC-L1)<sup>41,42</sup> and Principal Component Analysis (PCA)<sup>43</sup>. These methods have been previously used in various studies<sup>36,37,44–49</sup>. Subsequently, prediction models have been developed implementing several machine learning techniques like ExtraTrees<sup>50</sup>, KNN, Random forest<sup>51</sup>, Logistic Regression (LR)<sup>52</sup>, Ridge classifier<sup>53</sup> and SVC - RBF kernel with class weight factor employing scikit package<sup>54</sup> (described in Methods). The pipeline depicting the workflow of this study is shown in Fig. 1.

**Gene expression based models.** With an aim to classify the metastatic and primary tumor samples with high precision, first the RNAseq expression data of 466 patients consisting of 20,502 genes was used to select the relevant features using three feature selection methods; SVC-L1, WEKA-FCBF and PCA. Primarily, we

Classifiers	Dataset	TP	FP	TN	FN	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC
ETrees	Training	268	12	69	22	92.41	85.19	90.84	0.75	0.95
	Validation	67	5	16	7	90.54	76.19	87.37	0.65	0.94
KNN	Training	269	9	72	21	92.76	88.89	91.91	0.78	0.95
	Validation	66	5	16	8	89.19	76.19	86.32	0.62	0.93
RF	Training	260	8	73	30	89.66	90.12	89.76	0.74	0.96
	Validation	66	2	19	8	89.19	90.48	89.47	0.73	0.95
LR	Training	261	8	73	29	90	90.12	90.03	0.74	0.97
	Validation	65	2	19	9	87.84	90.48	88.42	0.71	0.95
RC	Training	262	9	72	28	90.34	88.89	90.03	0.74	0.96
	Validation	65	2	19	9	87.84	90.48	88.42	0.71	0.95
SVC-W	Training	269	8	73	21	92.76	90.12	92.18	0.79	0.97
	Validation	66	2	19	8	89.19	90.48	89.47	0.73	0.95

**Table 1.** Performance measures of 17 mRNA expression based features (selected by SVC-L1 feature selection method) on training and independent validation dataset to classify metastatic from primary tumor samples applying various machine-learning algorithms (classifiers). ETrees: Extra Trees Classifier; KNN: K-Nearest Neighbors Classifier; RF: Random Forest; LR: Logistic Regression; RC: Ridge Classifier; SVC-W: Support Vector Classification with weight factor; TP: True positive; FP: False Positive; TN: True Negative; FN: False Negative; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews Correlation Coefficient; AUROC: Area under the Receiver Operating Characteristic.

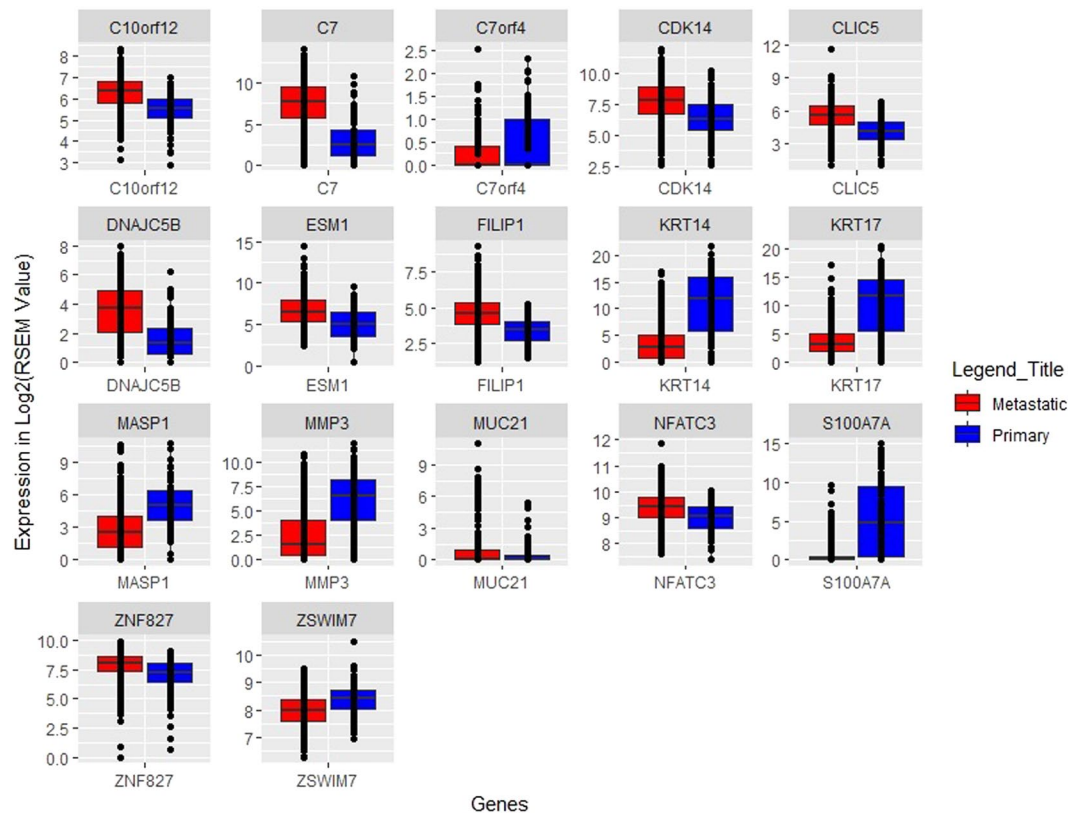
obtained nearly 150 and 17 features using WEKA-FCBF and SVC-L1, respectively. Further, we applied six different machine learning algorithms on the selected features obtained using the above three methods. As shown in Table 1, nearly 92.76% (sensitivity) metastatic tumors and 90.12% (specificity) primary tumors of training dataset and 89.19% (sensitivity) metastatic and 90.48% (specificity) primary tumor samples of validation dataset are correctly identified by SVC-model based on these 17 features (selected by SVC-L1). This model achieved accuracy of 92.18% and 89.47% with AUROC of 0.97 and 0.95 on training and validation dataset, respectively (Table 1). We have selected these above threshold dependent measures based on the threshold of SVM score (decision function in scikit) that gave maximum accuracy along with the minimum difference between sensitivity and specificity (Supplementary Table S1). The boxplot depicting the expression pattern of these 17 features in metastatic and primary tumor samples is shown in Fig. 2.

Interestingly, a model based on 150 features selected using WEKA-FCBF also attained almost similar performance (Supplementary Table S2). Further, we also selected 32 Principal Component features using Principal Component Analysis (PCA), each of which explains at least 1% of the variance in the data. Logistic Regression (LR) based prediction model performed best, classifying metastatic and primary samples with 92.96% sensitivity, 76.19% specificity, 89.13% accuracy with 0.91 AUROC on validation dataset (Supplementary Table S3). As the models based on features selected by SVC-L1 have smaller number of features and higher performance as compare to the models based on features selected by WEKA-FCBF and PCA, respectively. We considered and reported the model based on 17 features as best expression-based classification model to distinguish metastatic and primary tumor samples of SKCM.

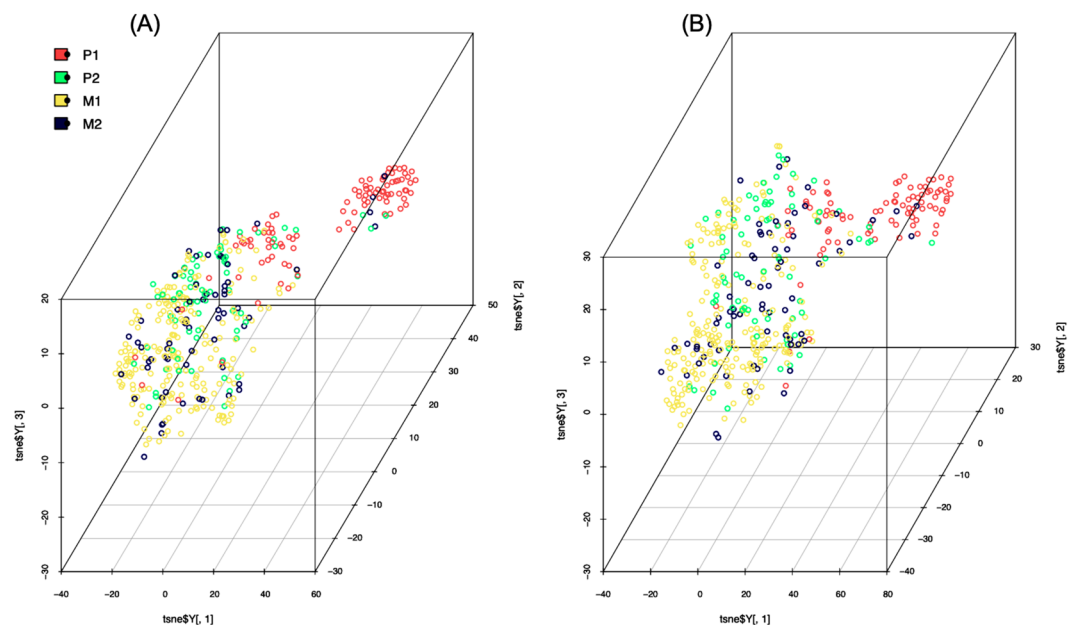
The enrichment analysis of the 17 features shows the biological role of the mRNA signature in melanoma carcinogenesis. Out of 17 genes, *C7* and *MASPI* are involved in Complement system activation (adjusted p-value < 0.05), while *KRT17* and *KRT14* are part of intermediate filament component (adjusted p-value < 0.05). It has been shown that metastatic cancer cells use actin bundles to disrupt from a primary tumour and invade the surrounding tissue. After travelling in the vasculature or lymphatic system, they exit into a new niche and form a new tumour<sup>55</sup>.

As we analysed the new tumour event (NTE) clinical file of SKCM patients, we observed that 16 patients with primary tumor have been shown to be in distant metastasis with new tumour events. Therefore, we removed these 16 samples from the dataset and again developed the classification model. There was a marginal increase of MCC from 0.73 to 0.77 and alike AUROC on validation dataset (Supplementary Table S4).

From the above analysis, we have observed that 17 mRNA expression-based features are performing reasonably well in classifying metastatic and primary tumor samples. Further, we visualised the samples based on 17 mRNA expression features using t-SNE (t-Distributed Stochastic Neighbour Embedding) implemented using the *Rtsne*<sup>56</sup> and *scatterplot3D*<sup>57</sup> packages in R. The substantial number of P2 samples differ from P1, but some of them merge/co-clustered with P1, which is quite expected as P1 progresses to P2 (Supplementary Fig. S1(A)). The t-SNE analysis shows a clear distinction between P1 and M1 (Supplementary Fig. S1(B)) with some of the primary samples going extreme into the boundaries of M1. Surprisingly the distant metastatic samples are quite widely distributed in comparison to Primary tumors as shown in Supplementary Fig. S1(C). Next, Supplementary Fig. S1(D) presents the P1 tumors in contrast to P2 from M1 tumors. Here P1 tumors looks separated from P2 and M1 whereas P2 and M1 tumors are amalgamated. The Fig. 3(A) shows that primary tumor (P1) samples form the separate cluster (red colour) in comparison to different states of metastasis for all the samples and Fig. 3(B) shows all the four classes after removing 16 primary samples of NTE. This analysis prompt us to further develop specific prediction models for classifying each state of metastasis from primary samples.

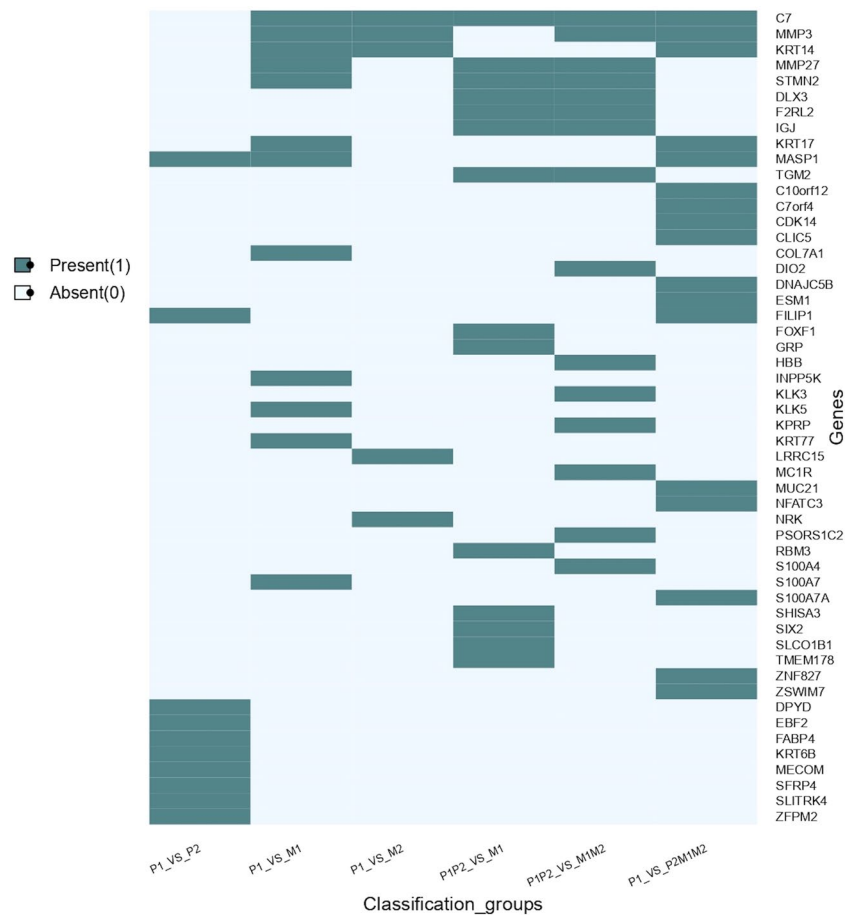


**Figure 2.** The expression pattern of 17 genes selected using SVC-L1.



**Figure 3.** The scatterplot3D view of tSNE dimension reduction of 17 selected features: (A) distribution of P1, P2, M1 and M2 samples; (B) distribution of P1, P2, M1 and M2 samples after removing 16 primary tumor samples (observed as distant metastatic in NTE file).

**Discrimination between Primary and sub-categories (or various states) of metastasis. Intra-lymphatic tumors v/s primary tumors.** The primary tumors (P1) are localised lesions and P2 includes the samples with-in-transit metastasis and satellite metastasis which represent intra-lymphatic tumour. At this stage, tumor has not still spread to lymphatic nodes. We selected 10 features using SVC-L1 (as in the above models,



**Figure 4.** The presence and absence of various features in different gene signatures developed for segregating metastatic samples from primary samples.

RNAseq data selected the appropriate number of features using SVC-L1) and the results of classification models on these 10 features show that it is difficult to classify the samples with intra-lymphatic tumour (P2) from the primary tumour (P1). The KNN-based model correctly identified 84.75% (Sensitivity) of metastatic and 93.83% (Specificity) of primary tumor patients of training data with MCC of 0.79 and AUROC of 0.96. On validation data, this model identified 73.33% (Sensitivity) of metastatic and 85.71% (Specificity) primary patients correctly with MCC 0.60 and AUROC of 0.84 (Supplementary Table S5). The selected ten features are shown in the Lane 1 of heatmap (Fig. 4).

**Lymphatic tumors v/s primary tumors.** Further, we tried to classify tumors that invaded lymphatic nodes (M1) from the primary tumors (P1). Our analysis shows that these tumors can be classified with high precision. The SVC-W based model using mRNA expression of 12 genes (Lane 2 of Fig. 4), selected using SVC-L1 feature selection method distinguished samples with good sensitivity of 97.74%, specificity 91.36% and MCC of 0.90 and AUROC of 0.98 on training data. We also observed the good sensitivity of 95.56% and specificity of 90.48% along with MCC of 0.86 and AUROC of 0.94 on the validation dataset. This indicates that once the tumour has reached the lymph nodes, there is substantial variation in the expression of genes associated with metastasis in comparison to the primary or localized tumor (Table 2).

**Distant metastatic tumors v/s primary tumors.** Next, we tried to classify the distant metastatic tumors (M2) from primary tumors (P1). Surprisingly the classification of these two groups of samples is not as good as lymphatic node v/s primary on 5 features (Lane 3 of Fig. 4). The KNN model correctly classified 87.04% (Sensitivity) distant metastatic samples and 92.59% (Specificity) primary samples with MCC of 0.80 and AUROC of 0.92 on training data. This model classified 78.57% (Sensitivity) distant metastatic samples and 85.71% (Specificity) primary samples correctly with MCC of 0.64 and AUROC of 0.81 on validation dataset (Supplementary Table S6).

**Regional v/s lymphatic tumors.** To differentiate between the tumors which have spread to lymph nodes (M1) and regional tumors, we combined primary (P1) and in transit and satellite tumors (P2). The LR model using 14 features (Lane 4 of Fig. 4) selected by SVC-L1, achieved the sensitivity of 92.09% and specificity of 90% with MCC of 0.82 and AUROC of 0.96 on training data and the sensitivity of 93.33% and specificity of 83.33% with MCC of 0.78 and AUROC of 0.89 on validation dataset (Supplementary Table S7).

Classifier	Dataset	TP	FP	TN	FN	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC
ETrees	Training	170	10	71	7	96.05	87.65	93.41	0.85	0.96
	Validation	44	3	18	1	97.78	85.71	93.94	0.86	0.91
KNN	Training	175	10	71	2	98.87	87.65	95.35	0.89	0.95
	Validation	43	4	17	2	95.56	80.95	90.91	0.79	0.92
RF	Training	155	7	74	22	87.57	91.36	88.76	0.76	0.96
	Validation	37	2	19	8	82.22	90.48	84.85	0.69	0.93
LR	Training	174	8	73	3	98.31	90.12	95.74	0.9	0.98
	Validation	43	2	19	2	95.56	90.48	93.94	0.86	0.93
RC	Training	175	10	71	2	98.87	87.65	95.35	0.89	0.97
	Validation	44	3	18	1	97.78	85.71	93.94	0.86	0.95
SVC-W	Training	173	7	74	4	97.74	91.36	95.74	0.9	0.98
	Validation	43	2	19	2	95.56	90.48	93.94	0.86	0.94

**Table 2.** Performance measures of 12 mRNA expression features (selected using SVC-L1 feature selection method) to discriminate M1 from P1 on training and independent validation dataset by applying various machine-learning algorithms. Etrees: Extra Trees Classifier; KNN: K-Nearest Neighbors Classifier; RF: Random Forest; LR: Logistic Regression; RC: Ridge Classifier; SVC-W: Support Vector Classification with weight factor; TP: True positive; FP: False Positive; TN: True Negative; FN: False Negative; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews Correlation Coefficient; AUROC: Area under the Receiver Operating Characteristic.

Classifier	Dataset	TP	FP	TN	FN	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC
ETrees	Training	250	14	62	28	89.93	81.58	88.14	0.67	0.92
	Validation	63	5	14	8	88.73	73.68	85.56	0.59	0.88
KNN	Training	234	12	64	44	84.17	84.21	84.18	0.61	0.91
	Validation	60	3	16	11	84.51	84.21	84.44	0.61	0.89
RF	Training	232	8	68	46	83.45	89.47	84.75	0.64	0.97
	Validation	63	3	16	8	88.73	84.21	87.78	0.67	0.95
LR	Training	241	14	62	37	86.69	81.58	85.59	0.62	0.93
	Validation	59	3	16	12	83.1	84.21	83.33	0.59	0.87
RC	Training	245	10	66	33	88.13	86.84	87.85	0.69	0.94
	Validation	62	4	15	9	87.32	78.95	85.56	0.61	0.89
SVC-W	Training	240	7	69	38	86.33	90.79	87.29	0.69	0.94
	Validation	64	4	15	7	90.14	78.95	87.78	0.66	0.89

**Table 3.** Performance measures of 32 miRNA expression features (selected by WEKA-FCBF feature selection method) on training and independent validation to classify metastatic from primary samples dataset by applying various machine-learning algorithms. Etrees: Extra Trees Classifier; KNN: K-Nearest Neighbors Classifier; RF: Random Forest; LR: Logistic Regression; RC: Ridge Classifier; SVC-W: Support Vector Classification with weight factor; TP: True positive; FP: False Positive; TN: True Negative; FN: False Negative; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews Correlation Coefficient; AUROC: Area under the Receiver Operating Characteristic

*Metastatic tumors v/s regional or primary tumor.* Further, we developed the model to categorize the tumors which spread to lymph nodes or metastasized (M1 and M2) from the tumors which were localized (P1 and P2). The Logistic Regression (LR) model based on 15 features (Lane 5 of Fig. 4) correctly classified 89.61% metastatic samples of training dataset with MCC of 0.74 and AUROC of 0.93. On validation dataset 81.36% metastatic sample and 80.56% of primary samples are correctly predicted with MCC of 0.61 and AUROC of 0.90 (Supplementary Table S8). The Lane 6 of Fig. 4 shows the 17 mRNA signature that has performed best out of all the combinations.

**miRNA expression based models.** Next, we have explored miRNA expression to elucidate its role in the progression of metastasis in SKCM. The number of miRNA features selected by WEKA-FCBF and SVC-L1 is 32 and 5 features, respectively. The SVC-W model based on 32 miRNAs attained the maximum performance with MCC of 0.69 and AUROC of 0.94 on training dataset and MCC of 0.66 and AUROC of 0.89 on the validation dataset. Further, nearly 86.33% (sensitivity) metastatic samples and 90.79% (specificity) primary samples of training dataset and 90.14% metastatic samples and 78.95% primary samples of validation dataset were correctly predicted (Table 3). The mean expression pattern of these 32 miRNA in primary and metastatic samples is represented in Supplementary Fig. S2. The Logistic Regression model based on the 5 miRNAs (feature selected by SVC-L1 method), achieved maximum MCC of 0.62 and AUROC of 0.93 on training dataset and MCC of 0.59 and AUROC of 0.87 on the validation dataset. This model correctly predicted 86.69% metastatic samples and 81.58% primary tumor samples of training dataset and 83.1% metastatic samples and 84.21% primary samples of

Classifier	Dataset	TP	FP	TN	FN	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC
ETrees	Training	227	18	58	51	81.65	76.32	80.51	0.52	0.88
	Validation	59	3	16	12	83.1	84.21	83.33	0.59	0.86
KNN	Training	226	15	61	52	81.29	80.26	81.07	0.54	0.88
	Validation	57	3	16	14	80.28	84.21	81.11	0.56	0.84
RF	Training	229	18	58	49	82.37	76.32	81.07	0.52	0.9
	Validation	56	3	16	15	78.87	84.21	80	0.54	0.88
LR	Training	241	14	62	37	86.69	81.58	85.59	0.62	0.93
	Validation	59	3	16	12	83.1	84.21	83.33	0.59	0.87
RC	Training	245	14	62	33	88.13	81.58	86.72	0.65	0.93
	Validation	58	3	16	13	81.69	84.21	82.22	0.58	0.88
SVC-W	Training	231	12	64	47	83.09	84.21	83.33	0.60	0.93
	Validation	54	3	16	17	76.06	84.21	77.78	0.51	0.87

**Table 4.** Performance measures of 5 miRNA expression features (selected by SVC-L1 feature selection method) on training and independent validation dataset to classify metastatic from primary samples by applying various machine-learning algorithms. Etrees: Extra Trees Classifier; KNN: K-Nearest Neighbors Classifier; RF: Random Forest; LR: Logistic Regression; RC: Ridge Classifier; SVC-W: Support Vector Classification with weight factor; TP: True positive; FP: False Positive; TN: True Negative; FN: False Negative; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews Correlation Coefficient; AUROC: Area under the Receiver Operating Characteristic.

validation dataset (Table 4). These 5 miRNAs include hsa-mir-205, hsa-mir-218.2, hsa-mir-513a.1, hsa-mir-675 and hsa-mir-7974. Here also, we filtered 43 Principal Component features employing Principal Component Analysis (PCA), each of them exhibits at least 1% variance of the data. The prediction model based on these features using Ridge Classifier method categorized metastatic and primary samples with an accuracy of 81.2% and 81.52% and AUROC 0.88 and AUROC 0.86 of training and validation datasets, respectively (Supplementary Table S9).

Among the miRNA signatures, hsa-mir-205 targets various genes (identified from miRTarBase<sup>58</sup>) such as *ZEB2*, *ZEB1*, *ERBB3*, *PRKC*, *ERBB2*, *E2F1*, *BCL2*, *ITGA5*, *VEGFA*, *AR*, *SMAD4*, *EGLN2*, *LAMC1*, *VEGFA*, *SMAD1*, *SRC*, *VEGFA*, *DDX5* and *YES1*, etc. Gene enrichment analysis have shown that these genes are significantly enriched (adjusted p-value < 0.05) in various cell growth promoting and oncogenesis associated pathways including transcriptional misregulation, TGF-beta signaling, wnt signaling, PDGF signaling, EGFR signaling, PI3K signaling, p53 signaling, ErbB signaling, VEGF signaling, cell cycle, hypoxia and angiogenesis, apoptosis processes, etc. This analysis signifies the role of hsa-mir-205 as tumor suppressor in melanoma development as it gets downregulated with the progression of metastatic melanoma.

**Methylation based model.** To ascertain the role of epigenetics in distinguishing metastatic from primary tumors, we took average methylation beta values for each gene as described in methods. Firstly, 38 and 2 features were selected using WEKA-FCBF and SVC-L1, respectively. Subsequently, classification models were developed using 38 features, and it can be observed in Table 5 that average methylation values are not very good predictors for distinguishing metastatic and primary tumor samples as compared to gene and miRNA expression. For instance, the LR model based on these 38 features achieved maximum performance, able to discriminate them with maximum MCC of 0.48 and 0.44 on training and validation dataset, respectively. It correctly predicted only 76.47% metastatic samples and 79.27% primary tumor samples of training dataset and 78.38% of metastatic samples and 71.43% primary tumor samples of validation dataset (Table 5). Further, 25 Principal Component features from methylation data filtered implementing PCA employing similar criteria like of mRNA and miRNA. SVC-W model based on these features is the best performer that attained an accuracy of 73.22% and 68.48% and AUROC 0.79 and AUROC 0.70 for segregation of tumor samples of training and validation datasets, respectively (Supplementary Table S10).

**Ensemble model.** Next, in order to compile information from individual models developed using all the three types of genomic features, we developed an ensemble method. In the ensemble method, prediction score from each model *i.e.* mRNA, miRNA and methylation were provided as input features to SVC. This model attained MCC of 0.73 along with AUROC of 0.97 and 0.71 MCC along with 0.93 AUROC on training and validation dataset, respectively (Table 6).

**Combo models.** As from the above analysis, we observe that 17 mRNAs and miRNA hsa-mir-205 have performed best for discriminating primary and metastatic tumours. Therefore, we combined them and developed models using various machine learning algorithms (Supplementary Table S11). The performance of this hybrid model is almost similar to the model based on 17 mRNA features with a marginal increase in specificity (Table 1).

Additionally, with an aim to extract information from all the three types of genomic features, *i.e.* RNAseq, miRNAseq and methylation-seq data, and develop multi-omics model, we combined all the features of three types of data by normalizing them using Min-Max normalization (see Methods). We used SVC-L1 method here as this feature selection method has shown reasonably higher performance with a minimum number of features

Classifier	Dataset	TP	FP	TN	FN	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC
ETrees	Training	248	17	65	41	85.81	79.27	84.37	0.6	0.89
	Validation	65	8	13	9	87.84	61.9	82.11	0.49	0.87
KNN	Training	224	19	63	65	77.51	76.83	77.36	0.47	0.83
	Validation	54	6	15	20	72.97	71.43	72.63	0.38	0.82
RF	Training	255	23	59	34	88.24	71.95	84.64	0.58	0.92
	Validation	65	9	12	9	87.84	57.14	81.05	0.45	0.87
LR	Training	221	17	65	68	76.47	79.27	77.09	0.48	0.84
	Validation	58	6	15	16	78.38	71.43	76.84	0.44	0.85
RC	Training	239	17	65	50	82.7	79.27	81.94	0.56	0.88
	Validation	62	8	13	12	83.78	61.9	78.95	0.43	0.83
SVC-W	Training	221	19	63	68	76.47	76.83	76.55	0.46	0.82
	Validation	58	6	15	16	78.38	71.43	76.84	0.44	0.91

**Table 5.** Performance measures of 38 features or average methylation of genes (features selected using WEKA-FCBF feature selection method) on training and independent validation dataset to classify metastatic from primary samples by applying various machine-learning algorithms. ETrees: Extra Trees Classifier; KNN: K-Nearest Neighbors Classifier; RF: Random Forest; LR: Logistic Regression; RC: Ridge Classifier; SVC-W: Support Vector Classification with weight factor; TP: True positive; FP: False Positive; TN: True Negative; FN: False Negative; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews Correlation Coefficient; AUROC: Area under the Receiver Operating Characteristic.

Dataset	TP	FP	TN	FN	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC
Training	244	5	71	34	87.77	93.42	88.98	0.73	0.97
Validation	60	1	18	10	85.71	94.74	87.64	0.71	0.93

**Table 6.** Performance measures of RNAseq, miRNAseq and methylation-seq ensemble features on training and independent validation dataset to classify metastatic from primary samples by applying SVC. ETrees: Extra Trees Classifier; KNN: K-Nearest Neighbors Classifier; RF: Random Forest; LR: Logistic Regression; RC: Ridge Classifier; SVC-W: Support Vector Classification with weight factor; TP: True positive; FP: False Positive; TN: True Negative; FN: False Negative; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews Correlation Coefficient; AUROC: Area under the Receiver Operating Characteristic.

in comparison to WEKA-FCBF and PCA in the previous analyses. The 20 features (Supplementary Table S12) selected by SVC-L1 method include 14 mRNA (genes), 1 miRNA and 5 methylation genes.

Subsequently, various prediction models developed based on these features employing different machine learning techniques. The performance of the most of the prediction models based on these features is in a similar range as of the performance of models based on 17 mRNA expression features for both on training and validation datasets, respectively (Supplementary Table S13).

**Single feature-based classification model using mRNA and miRNA expression.** Here, our goal is to develop single feature-based classification models that rank each gene and miRNA for its contribution to distinguish primary and metastatic tumor using threshold-based approach which was implemented in our previous studies for ranking of the genes<sup>36</sup>. In threshold-based model, a sample is classified as metastatic if the log<sub>2</sub> RNA-Seq by Expectation Maximization (RSEM) value of the feature (if feature is upregulated in metastatic) is higher than a threshold value, otherwise, it is classified as a primary sample. In these models, the threshold is varied incrementally from minimum to maximum RSEM value. Finally, that threshold is selected, which have the maximum AUROC in classifying metastatic and primary tumor samples. Consequently all the mRNA and miRNA sites are ranked on the basis of maximum AUROC and MCC with the minimum difference in sensitivity and specificity to assess the ability of each feature to classify metastatic and primary samples (Supplementary Table S14). Table S14 represents 20 mRNAs and 2 miRNA that can distinguish two types of samples with high precision.

The hsa-mir-205 and hsa-mir-203b are the top 2 miRNAs that can classify the metastatic and primary tumor samples with AUROC 0.83 and 0.75 at thresholds 4.3 and 1 (log<sub>2</sub> RSEM values), respectively. Both of these miRNAs are downregulated in metastatic samples, which indicates their potential role as tumor suppressors. The *C7*, *S100A7*, *LOC642587*, *CASP14* and *MMP3* are among the top 5 mRNA expression features that can discriminate metastatic and primary tumor samples with AUROC 0.81, 0.78, 0.77, 0.77 and 0.77 at thresholds 4.3, 3.1, 0.9, 0.9 and 3.7 (log<sub>2</sub> RSEM values), respectively. Among them, *C7* is upregulated in metastatic samples, while rest of the observed genes are downregulated in metastatic samples.

**Discriminating the early and late stage Primary SKCM tumors.** We further subdivided the heterogeneous P1 subgroup according to the SKCM tumor stage. Of the total 103 samples of primary SKCM, the tumor stage information is available for 98 patients with gene expression data and for 96 patients with miRNA expression data.



Classifier	TP	FP	TN	FN	Sen (%)	Spec (%)	Acc (%)	MCC	AUROC
Etrees	58	3	28	9	86.57	90.32	87.76	0.74	0.96
KNN	59	16	15	8	88.06	48.39	75.51	0.4	0.78
RF	64	5	26	3	95.52	83.87	91.84	0.81	0.96
LR	54	5	26	13	80.6	83.87	81.63	0.61	0.87
RC	53	8	23	14	79.1	74.19	77.55	0.51	0.83
SVC-W	62	9	22	5	92.54	70.97	85.71	0.66	0.88

**Table 7.** Performance measures of 37 mRNA expression features (selected using SVC-L1 feature selection method) to discriminate early from late stage primary tumors using leave one out cross validation by applying various machine-learning algorithms. Etrees: Extra Trees Classifier; KNN: K-Nearest Neighbors Classifier; RF: Random Forest; LR: Logistic Regression; RC: Ridge Classifier; SVC-W: Support Vector Classification with weight factor; TP: True positive; FP: False Positive; TN: True Negative; FN: False Negative; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews Correlation Coefficient; AUROC: Area under the Receiver Operating Characteristic.

Classifier	TP	FP	TN	FN	Sen(%)	Spec(%)	Acc(%)	MCC	AUROC
Etrees	52	3	27	9	85.25	90	86.81	0.72	0.93
KNN	56	2	28	5	91.8	93.33	92.31	0.83	0.96
RF	48	3	27	13	78.69	90	82.42	0.65	0.92
LR	55	0	30	6	90.16	100	93.41	0.87	0.99
RC	60	2	28	1	98.36	93.33	96.7	0.93	0.99
SVC-W	59	0	30	2	96.72	100	97.8	0.95	0.99

**Table 8.** Performance measures of 32 miRNA expression features (selected using SVC-L1 feature selection method) to discriminate early from late stage primary tumors using leave one out cross validation by applying various machine-learning algorithms. Etrees: Extra Trees Classifier; KNN: K-Nearest Neighbors Classifier; RF: Random Forest; LR: Logistic Regression; RC: Ridge Classifier; SVC-W: Support Vector Classification with weight factor; TP: True positive; FP: False Positive; TN: True Negative; FN: False Negative; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews Correlation Coefficient; AUROC: Area under the Receiver Operating Characteristic.

We have used both WEKA-FCBF and SVC-L1 based feature selection method (described in methods) to extract the important gene expression based features which could discriminate the early stage and late stage primary tumors. The WEKA-FCBF based method resulted in fewer features and better performance. Due to availability of a lesser number of samples (less than 100) we have used leave-one-out cross-validation technique to develop the prediction model using selected 37 (Supplementary Fig. S3) mRNA-based (WEKA-FCBF selected features) expression features. The random forest-based method performed best with a sensitivity of 95.52% and specificity of 83.87% with MCC of 0.81 (Table 7). Many of the genes in this signature have been already shown to be associated with melanoma.

One of the genes in the signature, *HSF1* has been already shown to be associated with early stage melanoma and has been shown to drive metastasis<sup>59</sup>. Another gene is *CDC37*, which is observed to be an essential gene to maintain the role of proteins that interact with protein kinases in melanoma. Furthermore, *RPS27* has been reported to have mutations in untranslated region and shown to have an impact in the progression of melanoma<sup>60</sup>. We could not find any of genes in this signature that is common with the genes that segregate primary and several forms of metastatic melanomas. This points out the heterogeneous nature of the primary melanomas itself.

Next, miRNA expression was explored to distinguish between the early stage and late stage primary SKCM samples. Using 32 miRNA (Supplementary Fig. S4) features selected by SVC-L1 feature selection method, KNN model is the top performer with balanced sensitivity of 91.8% and specificity of 93.33% with AUROC of 0.96 (Table 8). Of the 32 miRNAs, earlier few have been shown to be associated with melanoma, and many others have observed to be regulated in other cancers. For instance, hsa-mir-198 has been manifested to inhibit invasion of melanoma cells previously and has been downregulated in late stage as compared to an early stage in our analysis<sup>61</sup>. The expression of hsa-mir-219 has been shown to be downregulated in malignant melanoma (consistent with our analysis) and has also exhibited to be an important therapeutic target in melanoma<sup>62</sup>. Other miRNAs such as hsa-let-7f-1 has been implicated in lung cancer and renal cancer<sup>63,64</sup>, while hsa-mir-219a-1 in renal cancer<sup>65</sup>.

**Web server implementation.** To contribute the scientific community, we developed a web server, CancerSPP (Skin Cancer Progression Prediction). CancerSPP is designed for the prediction and analysis of metastatic and primary tumor of SKCM from RNAseq, miRNA and methylation expression data. The web server has two modules; Prediction module and Data analysis module.

**Prediction module.** This module permits the users to predict different states of metastatic samples and primary tumor samples *i.e.* Intra-lymphatic tumors ((P2) v/s Primary tumor (P1), Lymphatic tumors (M1) v/s Primary tumors (P1), Distant Metastatic tumors (M2) v/s Primary tumors (P1), Regional (P1P2) v/s Lymphatic tumors (M1) and Metastatic tumors (M1M2) v/s Regional tumor (P1P2) utilizing RSEM expression quantification values of signature genes. The user needs to submit the RSEM value of signature genes for every melanoma patient. In the input file, the number of patients represents the number of columns in the file. The output file contains the prediction outcome with a score. Greater the score, higher is the probability of correct prediction.

**Data analysis module.** This module is used to evaluate the role of each gene in various melanoma states such as regional metastatic, lymph node metastatic and distinct metastatic vs primary tumors based on mRNA and miRNA expression profiles. Moreover, it also incorporates threshold-based MCC of each feature and mean expression values for the RNAseq expression data in the primary and metastatic state of SKCM.

## Discussion and Summary

There is an emergence of synergized clinical and molecular profiles of cancer samples that can aid in predictive modelling for early tumor detection and progression. This prediction helps the physicians in making a suitable decision about the treatment course<sup>66,67</sup>. Previous studies concerning SKCM have focussed on determining its sub types<sup>9</sup> and survival<sup>27,31</sup>. Li *et al.* made an attempt to predict metastatic progression of melanoma tumor samples and predicted metastatic progression scores using mRNA and miRNA expressions individually; based on which they assigned primary and metastatic samples to primary and metastatic groups. Further, they also found a correlation between clinical characteristics of samples, *i.e.* Clark's level and lymph node status with metastatic progression score. Although all of metastatic samples were correctly assigned to the metastatic group; but, many of primary samples were incorrectly assigned to the metastatic group based on the metastatic progression score. They reported that the proportion of runs where a primary tumor specimen was classified as metastatic among the 10,000 runs was also highly non-uniform<sup>38</sup>. But, they did not report the performance of their method in terms of standard parameters (sensitivity, specificity, MCC, Accuracy, AUROC) as well as standard cross-validation techniques have not been implemented (*e.g.*, 10-fold CV, independent validation). Additionally, the classification models were not available to the public. Thus, the current study is an attempt to overcome these inadequacies.

The present study is an effort for the identification of genomic signatures that can classify both metastatic and primary samples with high precision based on mRNA, miRNA and methylation data. Further, our aim is to validate the performance of our prediction model on an independent dataset in addition to 10-fold internal cross validation. Additionally, we have also identified signatures that can further categorize different types of metastatic states, *i.e.* intra-lymphatic tumor, lymphatic tumor and distant metastatic tumor samples from the primary tumor samples. Furthermore, we also developed a web server to predict and analyse new data based on those identified markers.

In this study, we have identified discriminative genomic features using different feature selection methods and their classification prediction potential elucidated implementing various machine learning algorithms in the segregation of metastatic from primary tumor samples. Our analysis shows that the mRNA expression profile is the strongest predictor of metastasis as compared to miRNA expression and methylation profile. In the current study, the SVC-W model based on the expression of 17 mRNAs is the best performer in discriminating metastatic from primary tumors with overall accuracy of 89.47%, MCC of 0.73 and AUROC 0.95 on independent validation datasets (Table 1). Furthermore, the different models based on mRNA expression were also developed, which can differentiate primary tumors from several states of metastasis with high precision. Interestingly, it has been observed that primary tumor can be easily distinguished from tumors which have metastasized to lymph nodes; as compared to the tumors which have not still metastasized to lymph nodes. Many of the genes from our analysis panel have already been implicated in skin cancer. The *C7* gene has shown to be a potential tumor silencer gene and its expression is highly downregulated in various carcinomas such as ovarian cancer and non-small cell lung cancer (NSCLC)<sup>68</sup>. In current study, this gene alone can correctly predict 83.79% metastatic samples with MCC of 0.56 and AUROC of 0.81. Another gene *MMP3* has been reported to acts as melanoma suppressor gene<sup>69</sup> and is observed repeatedly in different signatures that distinguish various states of metastatic tumors from primary tumors in our analysis. Further, *KRT14*, a keratin gene, has shown to be downregulated in case of skin cancer<sup>70</sup>. In the present study, it classified metastatic and primary samples with high sensitivity of 94.14% and low specificity of 56.79% with overall the MCC of 0.56. Notably, the role of 11 out of 17 mRNA features have been previously reported in literature for cutaneous melanoma; while 6 genes including *ESM1*, *NFATC3*, *C7orf4*, *CDK14*, *ZNF827*, and *ZSWIM7* have been described for other cancers and other melanoma types like uveal melanoma but have not been specifically described for cutaneous melanoma<sup>71,72</sup>. Thus, the current study revealed the potential role of these six genes in the classification of the metastatic and primary tumor samples of SKCM for the first time. Earlier, the role of 8 out of 11 genes that include *C7*, *MMP3*, *KRT14*, *KRT17*, *MASP1*, *S100A7A*, *MUC21*, and *DNAJC5B* was previously reported in the metastatic progression of SKCM patients by Li *et al.*<sup>38</sup>. Martins *et al.* observed that one of the genes from our 17-gene signature, *i.e.* *C10orf12* gets upregulated with the loss of *ColVII* in squamous cell carcinoma model<sup>73</sup>. *CLIC5* which get upregulated in metastatic samples, has been previously shown to be methylated in one of 13 melanoma cell line<sup>74</sup>, while another study elucidated *FILIP1L* as a potential antivascular target for cancer therapy in melanoma model<sup>75</sup>.

Beside mRNA signatures, the miRNA and methylation features were also explored for segregation of primary and metastatic samples. Although these features did not segregate these samples as good as mRNA expression features, we were still able to find that expression of hsa-mir-205 is a strong predictor of metastatic melanoma. In our study, hsa-mir-205 alone can discriminates the metastatic and primary tumors with the sensitivity and specificity of 87.39% and 78.95%, respectively with MCC of 0.61 and AUROC of 0.83, if its expression is less than log<sub>2</sub> (RSEM value) of 4.3. The expression of hsa-mir-205 among the miRNAs has shown to be downregulated in

various solid tumors<sup>76</sup>. In the recent past, it has been observed that hsa-miR-205 targets oncogenes such as *E2F1* and *E2F5* and downregulates their expression which results in the inhibition of melanoma cell proliferation<sup>24</sup>. In addition, it also acts as tumor suppressor miRNA in skin carcinoma<sup>16,77</sup>, breast cancer<sup>78</sup> and prostate cancer<sup>79</sup>. We also used average methylation score to segregate the primary and metastatic samples and attained the sensitivity 78.38% and specificity 71.43% with MCC of 0.44 and AUROC of 0.91 with 38 features (Table 5). There is no single gene whose methylation score could segregate metastatic and primary samples well. Also the performance of this model based on 38 features is quite low as compared to the models obtained using gene expression and miRNA expression.

We also developed models combining different omic layers at the feature level (Combo model) and at the model level (ensemble model). Their performance was either less or comparable to the model based on 17 mRNA expression features. Further, we subdivided the heterogeneous group of primary tumors as per the tumor stage and segregated early stage and late stage samples. Based on 37 features, the random forest model segregated these samples with 95.52% sensitivity and 83.87% specificity with MCC of 0.81 and AUROC of 0.96 (Table 7). The features which segregate tumor stages (early and late) is different from features that segregate primary and metastatic samples.

Eventually, we assume that this study would be helpful to recognize important genomic signatures in the classification of primary tumor samples from the metastatic tumor samples of SKCM. Further, our analysis has shown that the genomic features selected by SVC-L1 feature selection method are fewer and have higher performance in classification of the SKCM samples into primary and metastatic classes as compare to the features selected by WEKA-FCBF and PCA methods, respectively. Thus, we hypothesized that this method might prove to be beneficial in scrutinizing important signatures from genomic data for diverse applications. Finally, we have developed the webserver CancerSPP to integrate all the prediction models and tools established in the current study. CancerSPP can analyze the gene expression data of a sample and predict whether it is a primary tumor or metastatic with a score using RSEM values derived from RNAseq and miRNAseq and methylation beta values.

## Material and Methods

**Datasets.** The RNAseq, miRNAseq and methylation profiling data for SKCM was retrieved from TCGA project using TCGA - Assembler 2 version<sup>80</sup>. In addition, manifest, biospecimen files and files containing clinical information such as new tumor events, drugs, age, gender, *etc.* were also downloaded to extract clinical parameters using Biospecimen Core Resource (BCR) IDs of patients/subjects. Finally, we obtained 466 patients [102 primary tumor, 74 Regional Cutaneous or Subcutaneous Tissue (includes satellite and in-transit metastasis), 222 Regional Lymph Node and 68 Distant Metastasis samples] for mRNA and methylation data, whereas 444 samples were available [95 primary tumor, 64 Regional Cutaneous or Subcutaneous tissue (includes satellite and in-transit metastasis), 214 Regional Lymph Node and 71 Distant Metastasis] for miRNA expression data. We referred primary tumors, Regional Cutaneous or Subcutaneous Tissue (includes satellite and in-transit metastasis), Regional Lymph Node and Distant Metastasis samples as P1, P2, M1 and M2, respectively. The clinical characteristics of these patients displayed in Supplementary Fig. S5.

In the present study, we have used mRNA and miRNA expression profiles in terms of RSEM values for 20,502 genes and 1,870 miRNAs, respectively. We downloaded the methylation profiles for 20,879 genes (acquired using the Illumina Human-Methylation450K DNA Analysis BeadChip assay, based on genotyping of bisulfite-converted genomic DNA at individual CpG-sites). Notable, we downloaded all CpG sites for each gene, (DNase hypersensitive and non-DNase hypersensitive). The data is in the form of Beta values, a quantitative measure of DNA<sup>81–83</sup>. Here, the methylation value for each gene represents the average of methylation beta values of all CpG sites located on each individual gene.

Further to study the Primary tumors stage-wise analysis using gene expression data, we segregated 67 stage-1 and stage-2 primary SKCM tumors as early stage and 31 stage-3 and stage-4 primary SKCM tumors as late stage SKCM tumors. In case of miRNA, 61 early stage and 30 late stage primary SKCM tumors are available for the analysis.

**Pre-processing of Data.** *Normalization of miRNA and mRNA expression.* Z-score Scaling: It has been observed that there is a wide range of variation in RSEM values of mRNAs and miRNAs. Thus, we transformed these values using log<sub>2</sub> after addition of 1.0 as a constant number to each of RSEM value. Further, features with low variance were excluded from the data using caret package in R<sup>84</sup>, followed by z-score normalization of data. Thus, log<sub>2</sub>-transformed RSEM values for each mRNA and miRNA were centred and scaled by employing caret package in R. Following equations were used for computing the transformation and normalization:

$$x = \log_2(RSEM + 1) \quad (1)$$

$$Z_{\text{score}} = \frac{x - \bar{x}}{sd} \quad (2)$$

Where  $Z_{\text{score}}$  is the normalized score,  $x$  is the log-transformed expression,  $\bar{x}$  is the mean of expression and  $sd$  is the standard deviation of expression.

Min-Max normalization: When we combined all the features from the three omics layers, the RNAseq expression, miRNA expression and methylation profiling data for feature selection in combo model, the Min-Max normalization method in R was employed using range option of preProcess from caret package. This ensured that all three types of features were in the same range of 0 and 1. The validation dataset was transformed in accordance with the training data using predict function in caret.

It was observed that in some of the patients, mRNA expression available for both tissue and blood samples. Here we took the average of both the samples for each patient.

**Feature selection techniques.** One of the challenges in developing the prediction model is to extract important features from the large dimension of features. Although, there are a number of methods for feature selection, we have used only those methods, which are well established and previously implemented in similar types of studies<sup>36,37,44–49</sup>. In this study, we implemented three techniques, *i.e.* SVC with L1 penalty employing Scikit package<sup>54</sup>, ‘SymmetricalUncertAttributeSetEval’ with search method of ‘FCBFSearch’ of WEKA software package<sup>85</sup> and PCA in R. We filtered genes (mRNA expression), methylation pattern of genes and miRNA expression as features that can distinguish metastatic samples from primary tumor samples using these techniques. The FCBF (Fast Correlation-Based Feature) algorithm employed correlation to identify relevant features in high-dimensional datasets in small feature space<sup>39</sup>. SVC-L1 method selects the non-zero coefficients and then applies L1 penalty to select relevant features to reduce dimensions of the data. For feature selection using PCA, we selected those Principal Components that represented at least 1% variance of the data.

To select the robust features, first, the data was split into the ratio of 80:20 for 10 times followed by features selection using SVC-L1 or WEKA-FCBF on each occasion from the training dataset. From this resampling process, we obtained 10 sub-sets of features. We have selected the subset having the highest performance. To check the robustness of features, we computed the average stability index (Jaccard index) using OmicsMarker package<sup>86</sup> for each subset and finally calculated the overall stability index. For all the signatures the average stability index is nearly in the range of 0.40 to 0.43.

**Implementation of machine learning techniques.** Firstly, we have developed the prediction models to categorize primary tumor and metastatic samples based on selected genomic features using various classifiers implementing Scikit package. These classifiers include ExtraTrees, KNN, Random forest, Logistic Regression (LR), Ridge classifier and SVC - RBF kernel with class weight factor were implemented employing scikit package. In addition, to understand the progression of skin cancer from primary tumor to metastasis, we also analyse and develop prediction models using various machine learning classifiers of scikit package based on genomic features (mRNA expression) to classify the sub-categories of metastatic samples from primary tumor samples *i.e.* Intra-lymphatic tumors (P2) v/s primary tumor (P1), lymphatic tumors (M1) v/s primary tumors (P1), distant metastatic tumors (M2) v/s primary tumors (P1), regional (P1P2) v/s lymphatic tumors (M1) and metastatic tumors (M1M2) v/s regional tumor (P1P2).

The optimization of the parameters for the various classifiers was done by using a grid search with area under PR (Precision Recall) curve as scoring performance measure for selecting the best parameter as our data was imbalanced.

**Visualization of samples.** After applying supervised learning to classify samples, we visualised the distribution of samples based on selected features on reduced dimensions using t-SNE methods implementing the two R Packages; Rtsne and scatterplot3d packages. t-SNE is a non-linear dimensionality reduction algorithm employed to analyze the high-dimensional data. It converts multi-dimensional data to two or more dimensions<sup>87</sup>.

**Identification of important features using simple threshold-based approach.** Here, we employed AUROC and MCC based feature selection technique to identify important features and developed single feature-based prediction models to distinguish metastatic samples from primary tumor samples. Single feature based models are also called threshold based models in which feature having a score below a specific threshold is assigned to metastatic tumor if it is downregulated in metastatic tumor samples otherwise it as primary tumor sample and vice versa. We computed performance of each given feature and identified features having the highest performance in terms of AUROC, MCC with minimum difference in sensitivity and specificity. Additionally, we have also computed their mean difference, log fold change, Bonferroni adjusted p-value using Wilcoxon test in R.

**Performance evaluation of models.** In the present study, both internal and independent validation techniques were employed to evaluate the performance of models. Previously, different studies employed the 80:20 ratio for the partitioning of a dataset into training and validation dataset<sup>36,37,88,89</sup>. Therefore, we implemented this standard protocol and subdivided our dataset into two subsets, *i.e.* training dataset and independent validation dataset in ratio of 80:20. We used 80% of the main dataset for training and remaining 20% for independent validation. First, the training dataset is used for developing model and for performing ten-fold cross-validation as internal validation. In this ten-fold-cross validation technique, training dataset is randomly split into ten sets; of which nine out of ten sets are used as training sets and the remaining tenth set as testing dataset. This process is repeated ten times in such a way that each set is exploited once for testing. The final performance of the trained model is the mean performance of all the ten sets.

In order to avoid the over-optimization of parameters in ten-fold cross-validation, we have also implemented independent validation. In the case of independent validation, we evaluated our model on an independent dataset, which was kept aside and remained unseen during feature selection and training or development of the model<sup>90</sup>.

In order to measure the performance of models, we used standard parameters. Both threshold-dependent and threshold-independent parameters were employed to measure the performance. In case of threshold-dependent parameters, we measured sensitivity, specificity, accuracy and Matthews correlation coefficient (MCC) using the following equations.

$$\text{Sensitivity (Sn)} = \frac{TP}{TP + FN} * 100 \quad (1)$$

$$\text{Specificity(Sp)} = \frac{TN}{TN + FP} * 100 \quad (2)$$

$$\text{Accuracy (Ac)} = \frac{TP + TN}{TP + FP + TN + FN} * 100 \quad (3)$$

$$\text{MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

Where, FP, FN, TP and TN are false positive, false negative, true positive and true negative predictions, respectively.

While, for threshold-independent measures, we used standard parameter AUROC. The AUROC curve is generated by plotting sensitivity or true positive rate against the false positive rate (1-specificity) at various thresholds. Finally, the area under ROC curve was calculated to compute a single parameter called AUROC. Notably, we have obtained classification performance in terms of sensitivity, specificity, accuracy, MCC, AUROC on various thresholds of prediction score for each of prediction models. We have selected only those threshold dependent measures based on the threshold of prediction score that gives maximum accuracy along with the minimum difference between sensitivity and specificity.

**Functional annotation of signature genomic markers.** In order to discern the biological relevance of the signature genes, enrichment analysis was performed using Enrichr<sup>91</sup>. Enrichr executes the Fisher exact test to identify enrichment score. It provides Z-score and the adjusted p-value which is derived by applying correction on the Fisher Exact test. Further, to understand the biological impact of miRNAs in metastatic melanoma development, we employed miRTarBase to identify target genes of signature miRNA.

Received: 9 October 2018; Accepted: 7 October 2019;

Published online: 31 October 2019

## References

- Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394–424, <https://doi.org/10.3322/caac.21492> (2018).
- Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA Cancer J Clin* **69**, 7–34, <https://doi.org/10.3322/caac.21551> (2019).
- Soura, E., Eliades, P. J., Shannon, K., Stratigos, A. J. & Tsao, H. Hereditary melanoma: Update on syndromes and management: Genetics of familial atypical multiple mole melanoma syndrome. *J Am Acad Dermatol* **74**, 395–407; quiz 408–310, <https://doi.org/10.1016/j.jaad.2015.08.038> (2016).
- Volkovova, K., Bilanicova, D., Bartonova, A., Letasiova, S. & Dusinska, M. Associations between environmental factors and incidence of cutaneous melanoma. Review. *Environ Health* **11**(Suppl 1), S12, <https://doi.org/10.1186/1476-069X-11-S1-S12> (2012).
- Ana-Teresa Maia, S.-J. S. Ana Jacinta-Fernandes. Big data in cancer genomics. *Current Opinion in Systems Biology* **4**, 78–84 (2017).
- Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**, 685–696, <https://doi.org/10.1038/nrg2841> (2010).
- Tomczak, K., Czerwinska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19**, A68–77, <https://doi.org/10.5114/wo.2014.47136> (2015).
- Garman, B. *et al.* Genetic and Genomic Characterization of 462 Melanoma Patient-Derived Xenografts, Tumor Biopsies, and Cell Lines. *Cell Rep* **21**, 1936–1952, <https://doi.org/10.1016/j.celrep.2017.10.052> (2017).
- Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696, <https://doi.org/10.1016/j.cell.2015.05.044> (2015).
- Guan, J., Gupta, R. & Filipp, F. V. Cancer systems biology of TCGA SKCM: efficient detection of genomic drivers in melanoma. *Sci Rep* **5**, 7857, <https://doi.org/10.1038/srep07857> (2015).
- Greenberg, E. S., Chong, K. K., Huynh, K. T., Tanaka, R. & Hoon, D. S. Epigenetic biomarkers in skin cancer. *Cancer Lett* **342**, 170–177, <https://doi.org/10.1016/j.canlet.2012.01.020> (2014).
- Mazar, J. *et al.* Epigenetic regulation of microRNA genes and the role of miR-34b in cell invasion and motility in human melanoma. *PLoS One* **6**, e24922, <https://doi.org/10.1371/journal.pone.0024922> (2011).
- Mazar, J., DeBlasio, D., Govindarajan, S. S., Zhang, S. & Perera, R. J. Epigenetic regulation of microRNA-375 and its role in melanoma development in humans. *FEBS Lett* **585**, 2467–2476, <https://doi.org/10.1016/j.febslet.2011.06.025> (2011).
- Kanamaru, H. *et al.* The circulating microRNA-221 level in patients with malignant melanoma as a new tumor marker. *J Dermatol Sci* **61**, 187–193, <https://doi.org/10.1016/j.jdermsci.2010.12.010> (2011).
- Nguyen, T. *et al.* Downregulation of microRNA-29c is associated with hypermethylation of tumor-related genes and disease outcome in cutaneous melanoma. *Epigenetics* **6**, 388–394 (2011).
- Philippidou, D. *et al.* Signatures of microRNAs and selected microRNA target genes in human melanoma. *Cancer Res* **70**, 4163–4173, <https://doi.org/10.1158/0008-5472.CAN-09-4512> (2010).
- Schinke, C. *et al.* Aberrant DNA methylation in malignant melanoma. *Melanoma Res* **20**, 253–265, <https://doi.org/10.1097/CMR.0b013e328338a35a> (2010).
- Goto, Y. *et al.* Aberrant fatty acid-binding protein-7 gene expression in cutaneous malignant melanoma. *J Invest Dermatol* **130**, 221–229, <https://doi.org/10.1038/jid.2009.195> (2010).
- Tanemura, A. *et al.* CpG island methylator phenotype predicts progression of malignant melanoma. *Clin Cancer Res* **15**, 1801–1807, <https://doi.org/10.1158/1078-0432.CCR-08-1361> (2009).
- Zheng, H. *et al.* Down-regulation of Rap1GAP via promoter hypermethylation promotes melanoma cell proliferation, survival, and migration. *Cancer Res* **69**, 449–457, <https://doi.org/10.1158/0008-5472.CAN-08-2399> (2009).
- Mori, T. *et al.* Estrogen receptor-alpha methylation predicts melanoma progression. *Cancer Res* **66**, 6692–6698, <https://doi.org/10.1158/0008-5472.CAN-06-0801> (2006).

22. Mori, T. *et al.* Predictive utility of circulating methylated DNA in serum of melanoma patients receiving biochemotherapy. *J Clin Oncol* **23**, 9351–9358, <https://doi.org/10.1200/JCO.2005.02.9876> (2005).
23. Hoon, D. S. *et al.* Profiling epigenetic inactivation of tumor suppressor genes in tumors and plasma from cutaneous melanoma patients. *Oncogene* **23**, 4014–4022, <https://doi.org/10.1038/sj.onc.1207505> (2004).
24. Dar, A. A. *et al.* miRNA-205 suppresses melanoma cell proliferation and induces senescence via regulation of E2F1 protein. *J Biol Chem* **286**, 16606–16614, <https://doi.org/10.1074/jbc.M111.227611> (2011).
25. Tiffen, J., Gallagher, S. J. & Hersey, P. EZH2: an emerging role in melanoma biology and strategies for targeted therapy. *Pigment Cell Melanoma Res* **28**, 21–30, <https://doi.org/10.1111/pcmr.12280> (2015).
26. Jensen, E. H. *et al.* Down-regulation of pro-apoptotic genes is an early event in the progression of malignant melanoma. *Ann Surg Oncol* **14**, 1416–1423, <https://doi.org/10.1245/s10434-006-9226-2> (2007).
27. Winnepenninckx, V. *et al.* Gene expression profiling of primary cutaneous melanoma and clinical outcome. *J Natl Cancer Inst* **98**, 472–482, <https://doi.org/10.1093/jnci/djj103> (2006).
28. Mischiati, C. *et al.* cDNA-array profiling of melanomas and paired melanocyte cultures. *J Cell Physiol* **207**, 697–705, <https://doi.org/10.1002/jcp.20610> (2006).
29. Smith, A. P., Hoek, K. & Becker, D. Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma *in situ* and advanced-stage melanomas. *Cancer Biol Ther* **4**, 1018–1029 (2005).
30. Talantov, D. *et al.* Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clin Cancer Res* **11**, 7234–7242, <https://doi.org/10.1158/1078-0432.CCR-05-0683> (2005).
31. Haqq, C. *et al.* The gene expression signatures of melanoma progression. *Proc Natl Acad Sci USA* **102**, 6092–6097, <https://doi.org/10.1073/pnas.0501564102> (2005).
32. Hoek, K. *et al.* Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. *Cancer Res* **64**, 5270–5282, <https://doi.org/10.1158/0008-5472.CAN-04-0731> (2004).
33. Geiger, T. R. & Peeper, D. S. Metastasis mechanisms. *Biochim Biophys Acta* **1796**, 293–308, <https://doi.org/10.1016/j.bbcan.2009.07.006> (2009).
34. Soong, S. J. *et al.* Predicting survival outcome of localized melanoma: an electronic prediction tool based on the AJCC Melanoma Database. *Ann Surg Oncol* **17**, 2006–2014, <https://doi.org/10.1245/s10434-010-1050-z> (2010).
35. White, R. R., Stanley, W. E., Johnson, J. L., Tyler, D. S. & Seigler, H. F. Long-term survival in 2,505 patients with melanoma with regional lymph node metastasis. *Ann Surg* **235**, 879–887 (2002).
36. Bhalla, S. *et al.* Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Sci Rep* **7**, 44997, <https://doi.org/10.1038/srep44997> (2017).
37. Jagga, Z. & Gupta, D. Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. *BMC Proc* **8**, S2, <https://doi.org/10.1186/1753-6561-8-S6-S2> (2014).
38. 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 Res* **28**, 453–463, <https://doi.org/10.1111/pcmr.12374> (2015).
39. Lei Yu, H. L. Feature selection for high-dimensional data: A fast correlation-based filter solution 856–863 (2003).
40. Frank, I. H. W. A. E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. 416 (Morgan Kaufmann Publishers, 1999).
41. and, H. H. & Sugiyama, M. Feature Selection via l1-Penalized Squared-Loss Mutual Information. *IEICE Transactions on Information and Systems* **E96-D**, 1513–1524 (2013).
42. Isabelle Guyon, J. W. A. S. B. In *Machine Learning* Vol. 46 389–422 (2002).
43. Hira, Z. M. & Gillies, D. F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv Bioinformatics* **2015**, 198363, <https://doi.org/10.1155/2015/198363> (2015).
44. Taguchi, Y. H. Principal Components Analysis Based Unsupervised Feature Extraction Applied to Gene Expression Analysis of Blood from Dengue Haemorrhagic Fever Patients. *Sci Rep* **7**, 44016, <https://doi.org/10.1038/srep44016> (2017).
45. Taguchi, Y. H., Iwade, M. & Umeyama, H. SFRP1 is a possible candidate for epigenetic therapy in non-small cell lung cancer. *BMC Med Genomics* **9**(Suppl 1), 28, <https://doi.org/10.1186/s12920-016-0196-3> (2016).
46. Taguchi, Y. H. Principal component analysis based unsupervised feature extraction applied to budding yeast temporally periodic gene expression. *BioData Min* **9**, 22, <https://doi.org/10.1186/s13040-016-0101-9> (2016).
47. Kamkar, I., Gupta, S. K., Phung, D. & Venkatesh, S. Stabilizing l1-norm prediction models by supervised feature grouping. *J Biomed Inform* **59**, 149–168, <https://doi.org/10.1016/j.jbi.2015.11.012> (2016).
48. Bastani, M. *et al.* A machine learned classifier that uses gene expression data to accurately predict estrogen receptor status. *PLoS One* **8**, e82144, <https://doi.org/10.1371/journal.pone.0082144> (2013).
49. Ma, S. & Huang, J. Penalized feature selection and classification in bioinformatics. *Brief Bioinform* **9**, 392–403, <https://doi.org/10.1093/bib/bbn027> (2008).
50. Wehenkel, P. G. A. E. Extremely randomized trees. **63**, 3–42 (2006).
51. Ho, T. K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 832–844 (1998).
52. Lin, H.-F. Y.-L. H.-J. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning* **85** (2011).
53. Kropf, M. G. a. S. In *From Data and Information Analysis to Knowledge Engineering* 684–691 (Springer, Berlin, Heidelberg, 2006).
54. Fabian Pedregosa, G. V., *et al.* Édouard Duchesnay Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* (2011) (2010).
55. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674, <https://doi.org/10.1016/j.cell.2011.02.013> (2011).
56. Maaten, L. v. d. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* **15**, 3221–3245 (2014).
57. Mächler, U. L. A. M. Scatterplot3d - an R Package for Visualizing Multivariate Data. *Journal of Statistical Software* **8**, 1–20 (2003).
58. Chou, C. H. *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* **46**, D296–D302, <https://doi.org/10.1093/nar/gkx1067> (2018).
59. Scott, K. L. *et al.* Proinvasion metastasis drivers in early-stage melanoma are oncogenes. *Cancer Cell* **20**, 92–103, <https://doi.org/10.1016/j.ccr.2011.05.025> (2011).
60. Dutton-Regester, K. *et al.* A highly recurrent RPS27 5'UTR mutation in melanoma. *Oncotarget* **5**, 2912–2917, <https://doi.org/10.18632/oncotarget.2048> (2014).
61. Weber, C. E. *et al.* miR-339-3p Is a Tumor Suppressor in Melanoma. *Cancer Res* **76**, 3562–3571, <https://doi.org/10.1158/0008-5472.CAN-15-2932> (2016).
62. Long, J., Menggen, Q., Wuren, Q., Shi, Q. & Pi, X. MiR-219-5p Inhibits the Growth and Metastasis of Malignant Melanoma by Targeting BCL-2. *Biomed Res Int* **2017**, 9032502, <https://doi.org/10.1155/2017/9032502> (2017).
63. Heinzelmann, J. *et al.* Specific miRNA signatures are associated with metastasis and poor prognosis in clear cell renal cell carcinoma. *World J Urol* **29**, 367–373, <https://doi.org/10.1007/s00345-010-0633-4> (2011).
64. Tong, A. W. Small RNAs and non-small cell lung cancer. *Curr Mol Med* **6**, 339–349 (2006).
65. Redova, M. *et al.* Circulating miR-378 and miR-451 in serum are potential biomarkers for renal cell carcinoma. *J Transl Med* **10**, 55, <https://doi.org/10.1186/1479-5876-10-55> (2012).

66. Cruz, J. A. & Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform* **2**, 59–77 (2007).
67. McCarthy, J. F. *et al.* Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. *Ann N Y Acad Sci* **1020**, 239–262, <https://doi.org/10.1196/annals.1310.020> (2004).
68. Ying, L. *et al.* Complement component 7 (C7), a potential tumor suppressor, is correlated with tumor progression and prognosis. *Oncotarget* **7**, 86536–86546, <https://doi.org/10.18632/oncotarget.13294> (2016).
69. McCawley, L. J., Wright, J., LaFleur, B. J., Crawford, H. C. & Matrisian, L. M. Keratinocyte expression of MMP3 enhances differentiation and prevents tumor establishment. *Am J Pathol* **173**, 1528–1539, <https://doi.org/10.2353/ajpath.2008.080132> (2008).
70. Alam, H., Sehgal, L., Kundu, S. T., Dalal, S. N. & Vaidya, M. M. Novel function of keratins 5 and 14 in proliferation and differentiation of stratified epithelial cells. *Mol Biol Cell* **22**, 4068–4078, <https://doi.org/10.1091/mbc.E10-08-0703> (2011).
71. Flockhart, R. J., Armstrong, J. L., Reynolds, N. J. & Lovat, P. E. NFAT signalling is a novel target of oncogenic BRAF in metastatic melanoma. *Br J Cancer* **101**, 1448–1455, <https://doi.org/10.1038/sj.bjc.6605277> (2009).
72. Delehedde, N. B. P. L. C. A. M. J. E. D. M. Vascular And Tumoral Expression Of Endocan / Esm-1 In Uveal Melanoma. *Investigative Ophthalmology & Visual Science* **53** (2012).
73. Martins, V. L. *et al.* Increased invasive behaviour in cutaneous squamous cell carcinoma with loss of basement-membrane type VII collagen. *J Cell Sci* **122**, 1788–1799, <https://doi.org/10.1242/jcs.042895> (2009).
74. Furuta, J. *et al.* Silencing of Peroxiredoxin 2 and aberrant methylation of 33 CpG islands in putative promoter regions in human malignant melanomas. *Cancer Res* **66**, 6080–6086, <https://doi.org/10.1158/0008-5472.CAN-06-0157> (2006).
75. Kwon, M. *et al.* Functional characterization of filamin A interacting protein 1-like, a novel candidate for antivascular cancer therapy. *Cancer Res* **68**, 7332–7341, <https://doi.org/10.1158/0008-5472.CAN-08-1087> (2008).
76. Hulf, T. *et al.* Discovery pipeline for epigenetically deregulated miRNAs in cancer: integration of primary miRNA transcription. *BMC Genomics* **12**, 54, <https://doi.org/10.1186/1471-2164-12-54> (2011).
77. Xu, Y., Brenn, T., Brown, E. R., Doherty, V. & Melton, D. W. Differential expression of microRNAs during melanoma progression: miR-200c, miR-205 and miR-211 are downregulated in melanoma and act as tumour suppressors. *Br J Cancer* **106**, 553–561, <https://doi.org/10.1038/bjc.2011.568> (2012).
78. Iorio, M. V. *et al.* microRNA-205 regulates HER3 in human breast cancer. *Cancer Res* **69**, 2195–2200, <https://doi.org/10.1158/0008-5472.CAN-08-2920> (2009).
79. Hulf, T. *et al.* Epigenetic-induced repression of microRNA-205 is associated with MED1 activation and a poorer prognosis in localized prostate cancer. *Oncogene* **32**, 2891–2899, <https://doi.org/10.1038/onc.2012.300> (2013).
80. Wei, L. *et al.* TCGA-Assembler 2: Software Pipeline for Retrieval and Processing of TCGA/CPTAC Data. *Bioinformatics*, <https://doi.org/10.1093/bioinformatics/btx812> (2017).
81. Bibikova, M. & Fan, J. B. GoldenGate assay for DNA methylation profiling. *Methods Mol Biol* **507**, 149–163, [https://doi.org/10.1007/978-1-59745-522-0\\_12](https://doi.org/10.1007/978-1-59745-522-0_12) (2009).
82. Bibikova, M. *et al.* High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* **16**, 383–393, <https://doi.org/10.1101/gr.4410706> (2006).
83. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587, <https://doi.org/10.1186/1471-2105-11-587> (2010).
84. Max K Contributions from Jed Wing, S. W., Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan. *Classification and Regression Training. R package version 6.0-71.*, <https://CRAN.R-project.org/package=caret> (2016).
85. Smith, T. C. & Frank, E. Introducing Machine Learning Concepts with WEKA. *Methods Mol Biol* **1418**, 353–378, [https://doi.org/10.1007/978-1-4939-3578-9\\_17](https://doi.org/10.1007/978-1-4939-3578-9_17) (2016).
86. Determan, C. E. Optimal Algorithm for Metabolomics Classification and Feature Selection varies by Dataset. *International Journal of Biology* **7**, <https://doi.org/10.5539/ijb.v7n1p100>. (2015).
87. Maaten, L. V. D. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* (2011) **15**, 3221–3245 (2014).
88. Agrawal, P. *et al.* In Silico Approach for Prediction of Antifungal Peptides. *Front Microbiol* **9**, 323, <https://doi.org/10.3389/fmicb.2018.00323> (2018).
89. Qureshi, A., Thakur, N. & Kumar, M. VIRsiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting human viruses. *J Transl Med* **11**, 305, <https://doi.org/10.1186/1479-5876-11-305> (2013).
90. Bhasin, M. & Raghava, G. P. SVM based method for predicting HLA-DRB1\*0401 binding peptides in an antigen sequence. *Bioinformatics* **20**, 421–423, <https://doi.org/10.1093/bioinformatics/btg424> (2004).
91. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90–97, <https://doi.org/10.1093/nar/gkw377> (2016).

## Acknowledgements

The authors acknowledge funding agencies J. C. Bose National Fellowship (DST). S.B., H.K., and A.D. are thankful to ICMR, CSIR and DST INSPIRE for providing fellowships.

## Author contributions

S.B. and H.K. collected the data and created the datasets, developed classification programs, implemented algorithms. S.B., H.K. and A.D. created the back-end server and front-end user interface. S.B., H.K., A.D. and G.P.S.R. analysed the results. S.B., H.K. and A.D. wrote the manuscript. G.P.S.R. conceived and coordinated the project, helped in the interpretation and analysis of data, refined the drafted manuscript and gave complete supervision to the project. All of the authors have read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-019-52134-4>.

**Correspondence** and requests for materials should be addressed to G.P.S.R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019