Predictive models for stage and risk classification in head and neck squamous cell carcinoma (HNSCC)

Sugandh Kumar[1,2], Srinivas Patnaik[2] and Anshuman Dixit[1]

[1] Computational Biology and Bioinformatics Laboratory, Institute of Life Science, Bhubaneswar, Odisha, India
[2] School of Biotechnology, Kalinga Institute of Industrial Technology (KIIT) University, Bhubaneswar, Odisha, India

ABSTRACT

Machine learning techniques are increasingly used in the analysis of high throughput genome sequencing data to better understand the disease process and design of therapeutic modalities. In the current study, we have applied state of the art machine learning (ML) algorithms (Random Forest (RF), Support Vector Machine Radial Kernel (svmR), Adaptive Boost (AdaBoost), averaged Neural Network (avNNet), and Gradient Boosting Machine (GBM)) to stratify the HNSCC patients in early and late clinical stages (TNM) and to predict the risk using miRNAs expression profiles. A six miRNA signature was identified that can stratify patients in the early and late stages. The mean accuracy, sensitivity, specificity, and area under the curve (AUC) was found to be 0.84, 0.87, 0.78, and 0.82, respectively indicating the robust performance of the generated model. The prognostic signature of eight miRNAs was identified using LASSO (least absolute shrinkage and selection operator) penalized regression. These miRNAs were found to be significantly associated with overall survival of the patients. The pathway and functional enrichment analysis of the identified biomarkers revealed their involvement in important cancer pathways such as GP6 signalling, Wnt signalling, p53 signalling, granulocyte adhesion, and dipedesis. To the best of our knowledge, this is the first such study and we hope that these signature miRNAs will be useful for the risk stratification of patients and the design of therapeutic modalities.

Subjects Bioinformatics, Computational Biology, Mathematical Biology, Oncology, Data Mining and Machine Learning
Keywords Head and neck cancer, TNM stage, Machine learning, Biomarker, microRNA, mRNA

Submitted 12 March 2020
Accepted 14 July 2020
Published 22 September 2020

Corresponding author
Anshuman Dixit,
anshumandixit@gmail.com,
anshuman@ils.res.in

Academic editor
Mitchell Stark

Additional Information and Declarations can be found on page 19

DOI **10.7717/peerj.9656**

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) is the 6th most common type of cancer (*Bray et al., 2018*) and is associated with 650,000 new cases and ~330,000 deaths annually worldwide (*Boscolo-Rizzo et al., 2018*; *Kapoor & Kumar, 2019*). The majority of the HNSCC cases are Oral Squamous Cell Carcinomas (OSCC) (*Deshpande & Wong, 2008*). More than 90% of HNSCC are associated with OSCC patients (*Vigneswaran & Williams, 2014*). The incidence rates (mainly OSCC (*Collaboration, 2019*)) are higher in South Asian countries such as India (*Poddar et al., 2019*), Bangladesh (*Collaboration, 2019*), and Pakistan (*Akhtar et al., 2016*) as compared to other parts of the world. There are several known risk factors of HNSCC such as chewing tobacco, smoking cigarettes,

excessive alcohol consumption (*Stenson, Brockstein & Ross, 2016*) and oncogenic virus such as Human papillomavirus (HPV) (*Marur et al., 2010*). Additionally, epigenetic regulation, mutation, copy number variation (CNV) and immune host response also play a key role in carcinogenesis (*Network, 2015*; *Leemans, Snijders & Brakenhoff, 2018*). Despite current advancement in cancer diagnosis and treatment, the overall 5-year survival rate is less than ∼50% in HNSCC due to a lack of proper diagnostic markers and targeted therapies (*Wiegand et al., 2015*).

It has been well documented that detection in early stages leads to higher survival as compared to late stages in various cancers including HNSCC. The American Joint Committee of Cancer staging (TNM) describes an early stage, primary tumor as ∼2–4 cm in diameter without lymph node proliferation and metastasis (TNM stage I and II). The tumor is considered advanced (late stage), if it is larger (>5 cm) and has spread into either nearby lymph nodes only (TNM stage III) or has metastasized to other parts of the body also (TNM stage IV) (https://www.cancer.org/treatment/understanding-your-diagnosis/staging.html).

A lot of new research have happened in HNSCC in the past few decades, however without clinically meaningful discoveries. While there are some biomarkers (e.g., HPV +ve and −ve), however, they lack important features, such as high specificity and sensitivity, low cost, and short turnaround time. A quick and accurate diagnosis would have numerous benefits for the patients such as proper treatment resulting in reduced morbidity as well as and improve treatment outcomes. Unfortunately, there is no such universally accepted biomarker for HNSCC accepted for clinical use. There is an urgent need for more effective therapies, and clinically relevant biomarkers to stratify patients in HNSCC.

The micro-RNAs (miRNAs) are ∼18–25 nucleotide long non-coding RNAs. They can regulate mRNA expression by interacting with the 3′ untranslated regions (UTR) leading to mRNA degradation. These miRNAs by virtue of their control over mRNA expression have important regulatory roles such as regulation of cell division, cell maturation, angiogenesis, proliferation, migration, invasion, metastasis, autophagy, and apoptosis (*MacFarlane & Murphy, 2010*). However, in various diseases especially cancers, these miRNAs themselves can get dysregulated leading to pathological conditions (*Esteller, 2011*). A large number of miRNAs have been quite well characterized for their biological function in cancer and their ability to regulate the expression of different cancer pathways (*Iorio & Croce, 2012*). It is also known that the changes in miRNA expression profile can be detected even before the appearance of clinical symptoms in some cancers (*Bianchi et al., 2011*). These miRNA due to their stability and ease of detection (in tissues as well as biological fluids) offer a rational approach for the development of excellent biomarkers (*Mostert et al., 2011*). The analysis of miRNA expression profiles may also offer an insight into underlying tumor progression and/or identifying new therapeutic targets.

Mathematical modeling has been widely used in disease modeling, classification and molecular function prediction for long. The advancement of next generation sequencing (NGS) technology and the availability of massive sequencing data have opened new avenues in disease process understanding and monitoring by machine learning. The machine learning techniques are generally used for risk stratification, mutational frequency

prediction, CNV, and new target identification. It has been suggested that machine learning (ML) techniques can be utilized for diagnosis and prognosis in cancers (*Obermeyer & Emanuel, 2016*). For instance, the ML techniques have been used for the detection of Rat Sarcoma (RAS) activation pathways in cancers utilizing expression data, SNPs (single nucleotide polymorphisms) and CNVs (copy number variations) (*Way et al., 2018*). A multi-parametric Decision Support System (DSS) with a multitude of heterogenic data (clinical information, genomics, and imaging data) has been used to predict the OSCC progression and potential relapse (local or metastatic) (*Komura & Ishikawa, 2018*). A study by *Avissar et al. (2009)* for identification of the miRNAs to predict the presence of HNSCC identified miR-221 and miR375 to be predictors. However, the study was done without considering stages/grade. Recently, a machine learning study used neural networks to predict recurrences in tongue cancer (*Alabi et al., 2019*). *Kim et al. (2019)* developed predictive models for survival prediction in oral cancer patients. There is no report to date that stratifies the clinical stages in HNSCC patients using miRNA expression profiles. The identification of late stages often involves cumbersome examination and invasive tests; therefore, such studies could be immensely useful in accomplishing better diagnosis and clinical outcomes. Such strategy can also be extended to the cancers that are difficult to reach and detect such as pancreatic cancer.

In the current study, we have analyzed the miRNA expression patterns in HNSCC patients in order to identify miRNA signatures that can distinguish stages (early and late stage) using machine learning approaches. It has to be remembered that the clinical stage is known to play an important role in the overall survival of the patients. These studies have led to the identification of signatures that can efficiently stratify the HNSCC patients into early and late clinical stages as well as to stratify patient's risk that may help in the decision of the treatment regimens e.g., whether it is effective to use less aggressive treatment than highly toxic therapies.

MATERIALS AND METHODS

The overall strategy involves the following steps. (1) data processing (2) data distribution into training (70%) and test set (30%) (3) model building (4) 10-fold repeated internal cross validation (5) independent external cross-validation. The detailed workflow has been provided in Fig. 1. The studies were divided into three parts (i) Identification of diagnostic signature: Involving the identification of signature miRNAs for patient's classification into early and late stage (ii) Identification of prognostic signature: evaluation of independent prognostic marker using L1-Cox Proportional Hazards Model with penalized regression to predict the risk to the patient (iii) Functional analysis: The identification of pathways and processes regulated/influenced by the identified biomarkers.

Data retrieval and pre-processing

The normalized reads of miRNAs (Illumina HiSeq 2000 platform, miRgene-level, Normalized), mRNAs (llumina HiSeq 2000 platform, Gene-level) expression were retrieved from the TCGA (https://cancergenome.nih.gov). A total of 528 samples (484 primary tumors and 44 Normal Adjacent Tissue (NAT)) with associated clinical information such

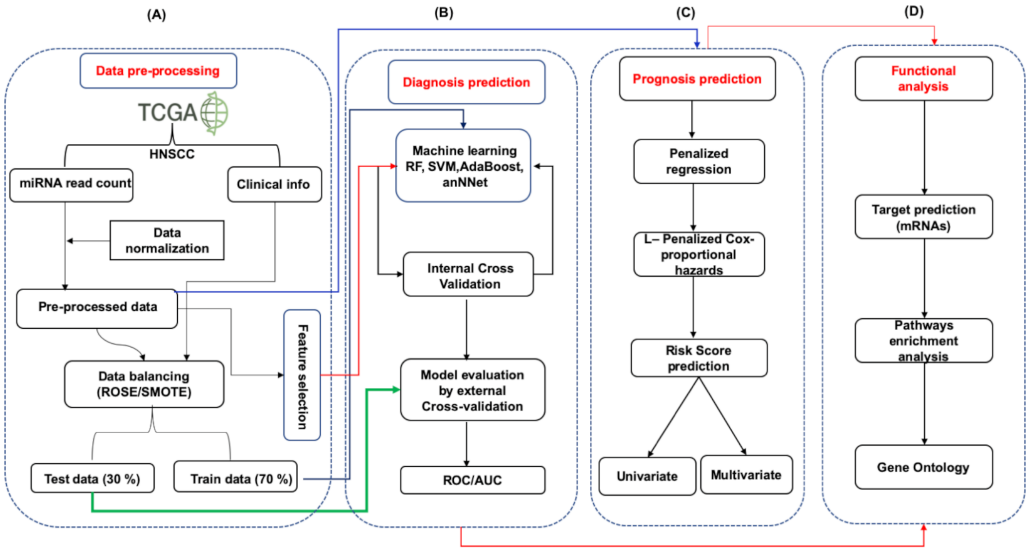

Figure 1 The schematic representation of workflow. The study was divided into four parts. (A) The expression data (mRNA and miRNA) along with clinical information of HNSCC patients was obtained from the TCGA and was preprocessed by Z-score normalization and balancing (ROSE/SMOTE) prior to model building. The data was then systematically divided into training (70%) and test (30%) sets. (B) Identification of signature miRNA for early and late stage classification was accomplished using the miRNA expression data. The training set was used for feature selection and model building using machine learning techniques viz. Random Forest (RF), Support Vector Machine using radial kernel (svmR), adaptive boosting (adaBoost), Average neural network (avNNet), and Gradient Boosting Machine (GBM). The test set was used as an external test set to rigorously and independently assess the performance of the generated model. (C) Prognostic miRNA signature was identified using the LASSO (least absolute shrinkage and selection operator) based L1-Cox-proportional hazards method. The individual miRNA was then accessed for their influence on the overall survival of the patients using the Kaplan-Meir test. (D) The functional analysis involved mRNA target prediction, pathways analysis, and Gene Ontology (GO).

Full-size DOI: 10.7717/peerj.9656/fig-1

as TNM stage, status, gender, and age etc. (Table S1) were selected whereas the samples with unknown stage information were removed. Finally, the data contained 27, 74, 78 and 274 samples from stage I–IV respectively (total 453 samples). The stages I and II were treated as early stage (total 101 samples) and stage III and IV as later stage (total 352 samples). Further, miRNAs and mRNAs expression profiles were normalized using the *limma-voom* (*Law et al., 2014*) library package in R3.3.5.

Differential miRNA and mRNAs expression analysis

Differential expression (DE) analysis of miRNAs and mRNAs were performed by the limma-voom using student's t-test. The mRNAs and miRNAs were considered differentially expressed if ($|\log_2 FC \geq 1|$ and p-value < 0.05). The volcano plot and heat map were generated using Enhancevolcano (*Blighe, 2018*) and Complexheatmap (*Gu, Eils & Schlesner, 2016*) library in *R3.3.5*.

Data scaling

In the current study the data was scaled using *Z-scaling*. The following equation was used for the scaling of data set.

$$x' = \frac{x - \overline{x}}{\sigma}. \tag{1}$$

Here, $x - \overline{x}$, represented the expression variance. While σ is the standard deviation of miRNAs expression. The x' is the normalized miRNAs expression. The '*Z-score*' method represent that data scaled to a mean of zero and a standard deviation of 1.

ML model building for stage classification

The "*caret*" library of R package (*Kuhn, 2012*) that contains several machine learning, complex regression and classification methods was utilized to develop various ML algorithm based classification models. The machine learning techniques were employed for building the models for patient classification into clinical stages using miRNA expression profiles. In this study, four widely used machine learning (ML) methods namely RF, svmR, AdaBoost, and avNNet were used.

The random forest (RF) method commonly uses non-linear regression models and has been applied in a variety of computational studies. It is a simple, interpretive, and flexible method which allows for a large number of predictor variables. Additionally, it also predicts well in the small sample sizes and high genetic heterogeneity (*Alexopoulos, 2010*). It comprises building trees from bootstrapped training data in such a way that each split is based on a random sample of m variables from a full set of n attributes (*Liaw & Wiener, 2002*).

$$IG(n) = 1 - \sum_{i=1}^{j} (pi)^2. \tag{2}$$

Here IG is Gini impurity of a node n is 1- the sum of overall predictor as j of the fraction of each pi square

Support Vector Machine (SVM) is popular ML technique in medical research for molecular function prediction, genomics variation, histopathology image classifications and subtyping of diseases (*Bianchi et al., 2011*). It is a non-probabilistic algorithm classifier based on the hyperplane line to maximize the maximum margin to separate two classes based on the two vectors points to achieve the best fit classification (*Huang et al., 2018*).

$$\int (x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i). \tag{3}$$

Here, function (x) is the kernel function which calculate the similarity between training and predicted x, x_i and α_i represent the parameter corresponding to each training and predictor variable. β_0 isthe constant.

The AdaBoost is an algorithm for producing strong classifiers from the weak classifier. It is based on the penalized weighted matrix of each instance and voting them according to their weight scheme. It is a popular machine learning method for both balance and

unbalanced data sets (*Zemel & Pitassi, 2001*). The AdaBoost equation is following

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right). \tag{4}$$

Where, $h_t(x)$ is the out of weak classifier of t for input x and α_t is weight given to model. The $\alpha_t = 0.5 * \ln(1-E)/E$. Here, E is based on the error rate of model.

The avNNeT is a kind of neural network capable of learning nearly infinite number of mapping functions. The neural network (NN) behaves like a natural human neuron. Every predictor variables (inputs) connects to an output response variable (output), either directly or through backward and forward propagation of single and several of hidden nodes to calculate units (neurons). The depth of an NN corresponds to the number of hidden layers. The calculation performed in NN are performed in a hidden layer called deep network (*Huang, Zhu & Siew, 2004*). The NN is commonly used for calculation of gene expression, CNV and clinical data for prognosis prediction of complex diseases such as cancer (*Mirza et al., 2019*).

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right). \tag{5}$$

Here, a_j^l of the jth neuron in the lth layer is related to the activation in the $(l-1)$th layer. The σ is vectorising function. The W^l is weight matrix of each layer, l. The jth row and kth column is w_{jk}^l . Also, for each layer l as define j a bias vector b_j^l.

The GBM algorithm is used to convert weak learners into strong learners. It is the tree-decision boosting based classification model where each of instances assigned an equal weight at the start. After the examination of first steps, then increase the weight of those observation that are difficult to classify and lower the weight for those that are easy to classify. This is the iterative process to optimized the best fit for classification (*Friedman, 2001*). The general equation used for the GBM is given below

$$h_m(x) = \sum_{j=1}^{j_m} b_{jm} 1R_{jm}(x). \tag{6}$$

Here, m-th step fit the decision tree hm (x). jm number leaves in each steps. The b_{jm} is the value predicted in the region of R_{jm}.

Data balancing: A challenge in biomedical studies is working with imbalanced data sets i.e., unequal number of normal and disease samples. Unbalanced predictive variable ratio do not meet the assumptions of the machine learning models and its predict biased. Therefore, SMOTE algorithm (*Chawla et al., 2002*) was used to balance the data using the 'DMwR' library in R.3.3.5 package. The SMOTE is a popular algorithm to deal with unbalanced data (*Lusa, 2013*; *Ramezankhani et al., 2016*). There are two types of sampling strategies commonly employed in machine learning viz. oversampling and under-sampling for unbiased classification. In under-sampling, the samples are reduced based on k nearest neighbor (kNN) clustering centroid distance while in oversampling minority classes are amplified to balance both predictors. In our data set, the number of late stage patient

samples was reduced so as to match the number of early stage patient samples ($N = 101$) (Fig. S1A) using the under-sampling method and in oversampling (Fig. S1B), early stage data was amplified ($N = 352$) to balanced early and later-stage samples.

Thereafter, the datasets (under and over sampled) were systematically distributed into training (70%) and test set (30%) separately. In under-sampled data, 141 and 61 samples were used as training and test set respectively whereas in oversampled data 478 and 206 samples were used as training and test set respectively. First, the model was trained using only training set the test set was kept aside for independent external cross-validation. A 10-fold repeated internal cross-validation with 10-time iterations to randomized data set was done to avoid generation of over/under fitting model. The cost functions were optimized (100 to 3,000 with 100 steps per iteration) to achieve accurate classification. The performance of the generated model was examined based on sensitivity, specificity, accuracy, and AUC (ROC), precision and MCC by external independent data. Equations are given below:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \tag{10}$$

Where TP is true positive; TN is true negative; FP false positive; and FN is false negative, MCC is Matthews correlation coefficient.

Patient's risk assessment based on signature miRNA expression

Further, we performed L-1 Penalized Estimation in the Cox Proportional Hazards Model using optimal cross-validated likelihood to identify prognostic markers based on the survival status (alive/dead), survival time and normalized miRNAs expression (*Acharya et al., 2017*). The Kaplan–Meier curve and the Log-rank method were used to estimate the significance of median miRNA expression on the survival of the patients. It evaluates the independent survival of the patient between the two risk groups (low/high). The patient's clinical parameters such as age, survival status, gender, metastasis, TNM stages, and median expression of significant miRNAs were evaluated for their association with patient's survival using cox-regression for a univariate and multivariate analysis.

The risk score for individual patients was calculated using L1-penalized (LASSO) likelihood Cox-regression (*Goeman, 2010*). As earlier, the data was partitioned into training (70%) and test (30%) to assess the performance of the generated models. The prognostic value of individual miRNAs was first calculated by the L1-penalised Cox univariate proportional hazard regression in R tool using a 'penalized' package. The significant miRNAs (p-value < 0.05) were used in subsequent Cox multivariate analysis. The prognostic model was built by the linear combination of the expression level of significant miRNAs multiplied with the Cox regression coefficient (β). The standard

formula was as follows:

$$\text{Risk Score} = X_1\beta_1 + X_2\beta_2 + \cdots + X_n\beta n + C. \quad (11)$$

Using the median risk score the patients were divided into high- and low-risk groups. The time-dependent receiver-operating characteristic (ROC) curve was created using "survivalROC" package in R to evaluate the performance (specificity and sensitivity) of the generated miRNA prognostic signature (*Chawla et al., 2002*).

The mRNAs target identification for identified miRNAs

The mRNAs targets of the identified biomarkers (diagnostic and prognostic) were predicted using the IPA (Ingenuity Pathway Analysis). The IPA predict miRNAs target based on three evidence (i) experimental (ii) high confidence (*Kozomara & Griffiths-Jones, 2014*) and (iii) low confidence. The experimental and high confidence interactions are curated by experts. In the current studies only experimentally validated and high confidence miRNA-mRNA interactions were used. The miRNAs can have a broad range of mRNAs targets and many of them may not be relevant to the disease under consideration. Therefore, the predicted target mRNAs were considered only if they are also significantly differentially expressed in the HNSCC. As a result, 2 sets (set 1 with 230 and set 2 with 262 mRNAs) were finally selected for diagnosis and prognostic miRNA targets. The miRNA-mRNAs interactions were further visualized for better understanding using Cytoscape3.7 (*Shannon et al., 2003*).

Pathway enrichment analysis

The previously identified genes (230) of set and 262 were further used for the identification of pathways and other molecular functions, network and toxicity enrichment analysis in the IPA (Ingenuity Pathway analysis). The IPA can analyze enriched biological pathways, disease and functions, upstream and downstream interactions, molecular function and toxicity for a given gene set. It also gives the statistical significance for each enriched process.

Gene ontology (GO) enrichment

The Gene Ontology is a vocabulary developed to compare and identify genes based on the processes in which they are involved. The GO is defined into three categories viz. Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). In the current studies, the enrichment analysis was performed using PANTHER (*Mi et al., 2019*) (http://www.pantherdb.org/geneListAnalysis.do) and enrichR (https://maayanlab.cloud/Enrichr/) with p-value < 0.001 for identification of significant GO terms.

RESULTS

One of the goals of this study was to develop an efficient miRNA signature to correctly classify the early and late stage patients using miRNAs expression profiles using the machine learning algorithms. Further, a least absolute shrinkage and selection operator (LASSO) penalized regression model was employed to identify signature miRNAs for their prognostic potential. The network and pathways analysis revealed that several target genes involving various tumor suppressor and oncogenes in cancer regulatory pathways may be affected by the aberrant expression of the identified miRNAs. The important results are given below.

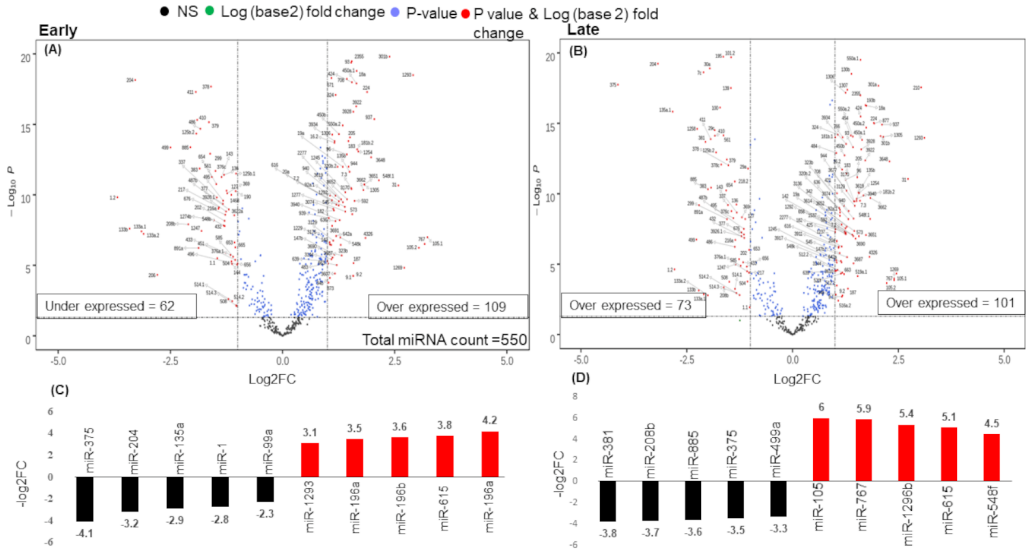

Figure 2 Volcano plot for expressed genes in (A) early stage (B) late stage. The miRNAs were considered significantly differentially expressed (DE-miRNA) if (|log2FC ≥ 1| and p-value < 0.05). Total 171 DE-miRNAs (under-expressed = 62 and over-expressed = 109) were identified in early stage whereas 174 DE-miRNAs (under-expressed = 73 and over-expressed = 101) were identified in the late stage. (C) Top five up (red) and down (black) regulated miRNAs in early stage (D) Top five up (red) and down (black) regulated miRNAs in late stage.

Full-size DOI: 10.7717/peerj.9656/fig-2

Differentially expressed microRNAs (DE-miRNAs)

There were 171 miRNAs differentially expressed (62 over and 109 under-expressed) in early and 174 miRNAs differentially expressed (73 over and 101 under-expressed) in late stages (Figs. 2A & 2B) (Tables S2A & S2B). The top 5 over and under-expressed miRNAs are given in Figs. 2C and 2D. The heat map of differentially expressed miRNAs in early stage shows a visual distinction between the expression of miRNAs in normal and cancer tissues (Figs. S2A & S3B).

An analysis of the common and stage specific miRNAs revealed that there are 148 miRNAs common to early and late stages, whereas 23 and 26 miRNAs are specific to early and late stages respectively (Fig. 3A) (Table S3). Similar to the DE-miRNAs analysis, the DE-mRNAs of 101 early stage with 44 normal adjacent tissues (NAT) samples were resulted in the 3831 differentially expressed mRNAs (2407 over and 1424 under-expressed) with (|log2FC ≥ 1| and p-value < 0.05) (Fig. S3).

Identification of miRNA signature for classification of early and late stages

An important objective of this study was to identify the miRNAs signature for the early and late (TNM stage) using the miRNAs expression profiles. The signature miRNAs were identified using the ensemble machine learning feature selection method (Random Forest (RF), Support Vector Machine Radial Kernel (svmR), Adaptive Boost (AdaBoost), averaged Neural Network (avNNet), and Gradient Boosting Machine (GBM)). A detailed method for feature selection by minimum Redundancy Maximum Relevance (mRMR) algorithm

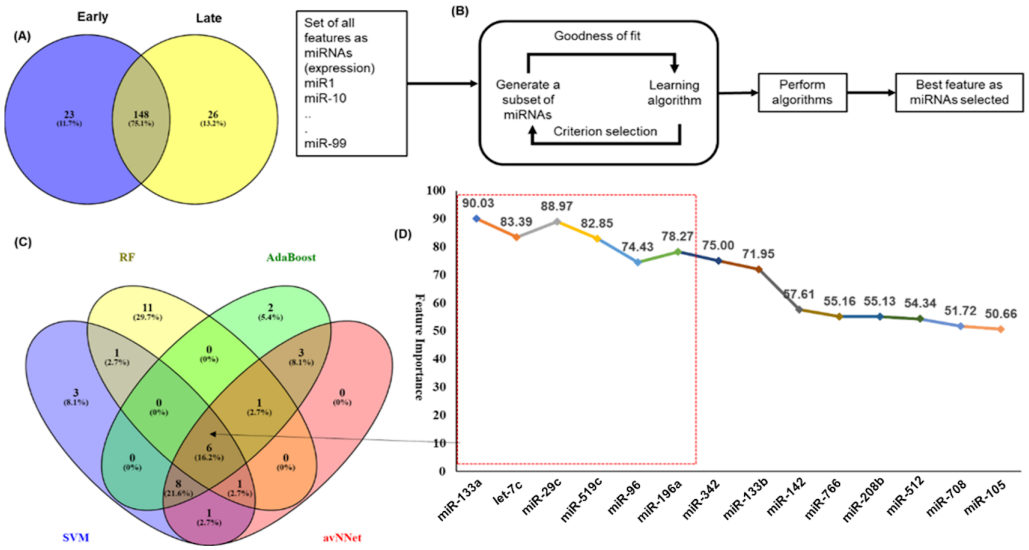

Figure 3 **Feature selection for model building.** The distribution of miRNAs into early and late stages. There were 23 miRNAs specific to early-stage while 26 miRNAs were specific to the late-stage. A total of 148 miRNAs were common to both the stages. (B) The schematic presentation of the feature selection in machine learning. (C) The common and specific features among top 20 features predicted by each of the five ML techniques. (D) The six common miRNAs (highlighted in red color box) predicted by all five ML methods and eight common in any three methods with the average importance for predictors.

Full-size DOI: 10.7717/peerj.9656/fig-3

is given in Fig. 3B. The top 20 predictor miRNAs were identified using each of the machine learning methods (RF, svmR, AdaBoost, avNNet and GBM) and the common among them were chosen for further model building and evaluation (Fig. 3C, Table S4). The ensemble ML based mRMR feature selection method reduce the bias that might be introduced by one or two methods. As a result, a 6 miRNA (let-7c, miR29c, miR-96, miR-133a, miR-196a, miR-519a) with average importance ranging from 90.03 to 78.27 for classification were selected for further studies. The mean importance of these miRNAs is shown in Fig. 3D. Among them, the miR-96 and miR-196a are over-expressed while the rest (miR-133b, miR-29c, miR-519a and let-7c) are under-expressed in normal vs primary tumor samples (Fig. S4). It is interesting to note that miR-519a is significantly differentially expressed in early stage only.

The performance of generated models for stage classification

The 6 miRNAs were selected by all five machine learning methods. The model was built using a training set and an independent test set (70:30) ratio was used for model prediction. The Synthetic Minority Oversampling Technique (SMOTE) algorithm was used for data balancing. In the under sampling total 101 (early and late stage) samples were used for internal model building and performance was evaluated using accuracy, sensitivity specificity, area under the curve (AUC), precision call and MCC of RF, svmR, AdaBoost, avNNet, and GBM. The accuracy of these five model was found to be 0.79 ± 0.03, 0.71 ± 0.04, 0.76 ± 0.03, 0.76 ± 0.03 0.79 ± 0.03, 0.71 ± 0.04, 0.76 ± 0.03, 0.76 ± 0.03, and 0.64 ± 0.042 for under-sampling methods. Similarly, the accuracy for RF, svmR, AdaBoost,

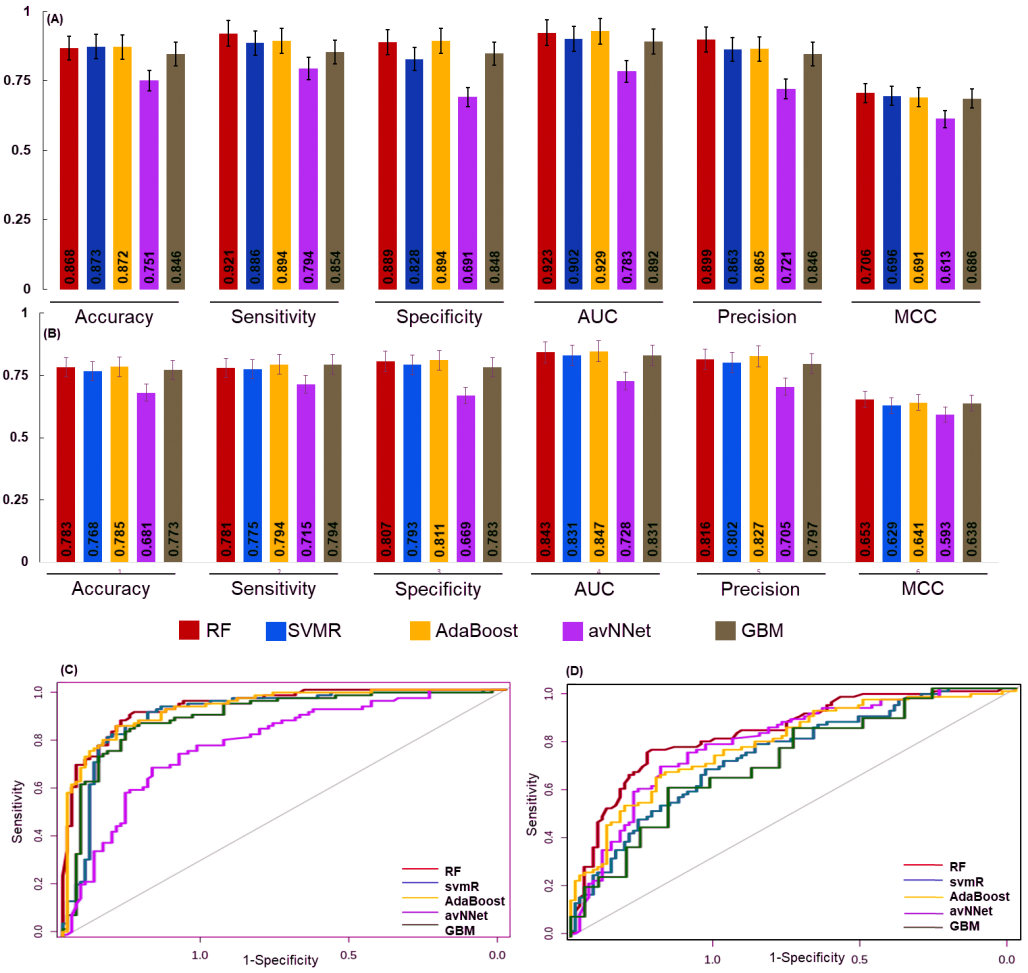

Figure 4 **The performance of the generated model.** The performance of the models was examined for accuracy, specificity, sensitivity and ROC generated for (A) over sampled and (B) under sampled data set. The model robustness was further evaluated by external data (test set) and AUC was shown in (C) over-sampling and (D) under sampling. In both the sampling methods the generated models showed good classification accuracy. The estimated error bars shown as standard error.

Full-size DOI: 10.7717/peerj.9656/fig-4

avNNet and GBM was 0.87 ± 0.021, 0.86 ± 0.032, 0.87 ± 0.035, 0.85 ± 0.032, and 0.80 ± 0.038 respectively for oversampled data (early and late = 352 samples) The details of the internal top 10 models of ML is given in the Table S5. The detailed performance of each ML algorithm is given in Figs. 4A & 4B.

The final model performance was evaluated by the independent test set using accuracy, sensitivity, specificity, AUC, precision, and MCC. The accuracy for RF, svmR, AdaBoost, avNNet, GBM was found 0.87, 0.87, 0.87, 0.75, and 0.84 for test set generated using oversampled data. Whereas the accuracy was 0.79, 0.76 0.78, 0.68, 0.77 for the above five methods for the test set generated using under-sampled data . The MCC (0.65, 0.62, 0.64, 0.59, and 0.68 was for RF, svmR, AdaBoost, avNNet and GBM) gave better model performance over other confusion matrix categories. The robustness of model was

evaluated using the area under curve (AUC) of Receiver Operating Characteristics (ROC) given in Figs. 4C & 4D for over and under sampled data. The RF and AdaBoost found the subsequently higher prediction accuracy in the under and over sampling methods and avNNet was the least prediction accuracy. The overall classification result showed, the generated 6 miRNAs signature can stratify the HNSCC patients on clinical TNM stages with reasonable accuracy based on expression profile.

Risk score calculation

To assess the prognostic performance of the generated model the time-dependent ROC curve was considered for risk score (low and high) calculation. The overall 5-year risk score was calculated of the normalized expression with the coefficient of the LASSO cox-coefficient. As stated in methodology, the risk score can be used to assess the risk for a patient based on the median value expression. The risk score was calculated using the risk score equation as given below and plotted using Kaplan-Meir plot Fig. 5. Kaplan-Meir plot for each of the signature miRNAs is given in Fig. S5. Risk Score $= (0.022 \times \text{let}-7_{\text{expression}}) + (-1.996 \times \text{miR-96}_{\text{expression}}) + (6.593 \times \text{miR-9}_{\text{expression}}) + (1.361 \times \text{miR-143}_{\text{expression}}) + (-2.938 \times \text{miR-379}_{\text{expression}}) + (-1.249 \times \text{miR-545}_{\text{expression}}) + (-1.969 \times \text{miR-658}_{\text{expression}}) + (-1.446 \times \text{miR-3926}_{\text{expression}}) - 1.505$.

The patient samples were systematically divided into training (368 samples) and test sets (160 samples). The 5 risk score (high and low) prediction was evaluated by AUC. The AUC for the training set and test set was found to be ~0.86 and ~0.79, respectively which indicated good performance of the identified signature (Figs. 6A & 6B). This time-dependent risk prediction analysis clearly showed that these miRNAs can be used for risk assessment of HNSCC patients.

Univariate and multivariate analysis for prognostic evaluation

The identified miRNA signature was evaluated against various clinical variables to check their effect on survival using the standard Cox-proportional hazards regression (Table 1). It indicates that advancing pathological stages may have a significant adverse impact on the survival of the patients. The Cox univariate analysis showed that pathological stage, T stage, N stage, metastasis and signature miRNAs were significantly associated with the prognosis. While in multivariate analysis pathological T stage (p-value $= 0.017$) and signature mRNA (p-value $= 0.023$) showed significant p-values associated with the poor prognosis (Table S6).

The microRNA-mRNA interactions

It was imperative to understand the role of the identified miRNA signatures in different biological processes. A mapping of the mRNA targets of the identified signatures (6 miRNAs for stage classification and 8 for prognosis) resulted in the identification of 1532 and 1446 mRNAs, respectively. Further, the analysis was done to select those mRNAs that get differentially expressed in various stages of HNSCC. It was found that for the identified diagnostic and prognostic miRNA signatures, there are 2 set of mRNA' target (set 1 with 230 and set 2 with 262) mRNA that were differentially expressed respectively in the HNSCC patients (Tables S7 & S8).

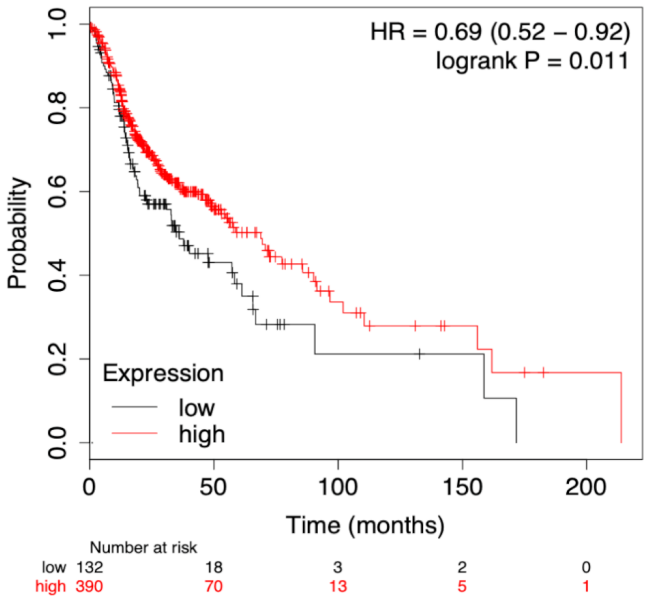

Figure 5 **Survival plot of prognostic signature miRNA as biomarker.** The cumulative effect of identified miRNA (eight in total) expression on the patient's survival. The cox-proportional hazard analysis revealed that the miRNAs were significantly associated with the overall survival of the patients (p-value < 0.05).
Full-size DOI: 10.7717/peerj.9656/fig-5

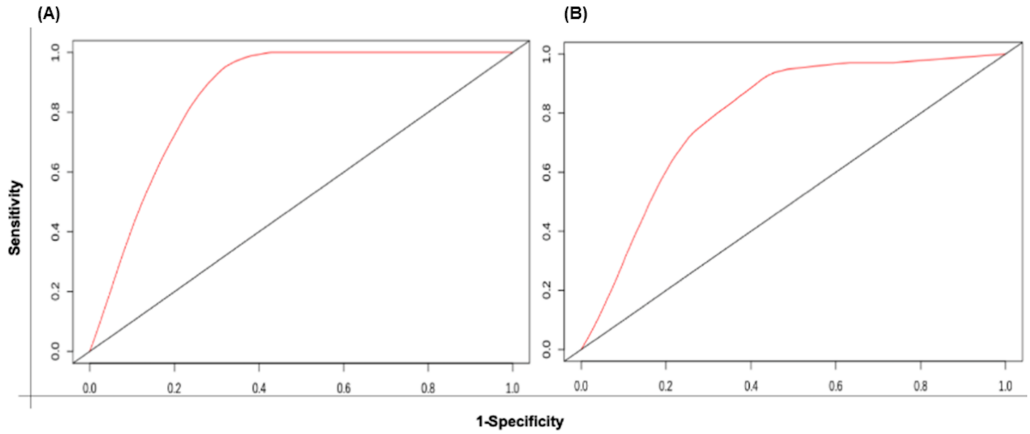

Figure 6 **The time-dependent ROC curve for evaluation of the prognostic performance of the generated miRNA signature for (A) training set and (B) test set.** The 5 years risk score prediction showed AUC of 0.86 in training set and 0.79 in the test.
Full-size DOI: 10.7717/peerj.9656/fig-6

Biological pathways, diseases functions and network enrichment analysis

As indicated earlier, the pathways enrichment analysis was performed by the Ingenuity Pathway Analysis (IPA) separately for diagnostic targets (set 1) and prognostic targets (set 2).

Table 1 The univariate and multivariate Cox regression analysis revealed independent risk factors.

Variables	Cox univariate analysis		Multivariate analysis	
	p-value	HR (95% CI)	*p*-value	HR (95% CI)
Age (high vs. low)	0.54	1.3 (0.99–3.18)		
Pathological stage (I + II vs. III + IV)	0.0082*	1.7 (1.1–2.40)		
Pathological T stage (T1 + T2 vs. T3 + T4)	3.1e−05*	2 (1.5–2.8)	0.0175*	3.4 (1.24–9.42)
Pathological N stage (N0 vs. N1 + N2)	0.00075*	1.8 (1.3–2.6)		
Pathological metastasis (M0 vs. M1)	0.0021*	28 (3.3–230)		
Gender (Male vs. Female)	0.35	0.72 (0.54–0.98)		
Signature miRNAs risk (Low vs. High)	0.044*	1.4 (1–2)	0.0235*	1.68 (0.88–3.20)

Analysis of set 1: The top enriched pathways for diagnostic targets were (i) GP6 signalling (ii) Neuroinflammation signalling (iii) Intrinsic Prothrombin Activation Pathway (iv) granulocyte adhesion and diapedisis (Fig. 7). Other enriched pathways are given in Table S9. The KEGG pathways analysis also revealed that majority of the genes were associated with pathways such as cytokine-cytokine receptor interaction, small cell lung cancer, transcriptional misregulation in cancer, ECM-receptor, and glutathione metabolism (Table S10). While WikiPathways enrichment showed, nuclear receptors meta-pathways (ID: WP2882), TGF-beta signalling pathways (WP366), PI3K-Akt-mTOR-signalling pathway (WP4172) (Table S11). The pathway analysis revealed that these miRNAs targets are highly enriched with cancer associated pathways. The top disease and disorders associated with the set 1 are organismal injury and abnormalities, cancer, connective tissue disorder, and skeletal muscle disorder. Top molecular functions and cellular functions were linked with cellular movement, cellular assembly and organization, cellular function and maintenance, cell death & survival and cell development (Fig. 8) (Table S12). The top upstream regulators were miR-29c, TGFB1, estrogen receptor, histone h4, and CREB which are associated with cell growth and proliferation. The miR-29, a potent tumor suppressor, is found under-expressed in this study. Interestingly, the top analysis of the upstream regulatory genes showed that miR-29 directly targets 11 genes (9 over-expressed and 2 under-expressed) associated with head and neck cancer (Fig. S6). A thorough look at these mRNAs revealed that many of them are transcription factors (total 31) and part of cancer gene panels (total 19) (*Sondka et al., 2018*). These mRNAs were used in the subsequent functional analysis step. It was imperative to note that many of the transcription factors and cancer censes gene are targeted by the let-7c which is well known for its role in a variety of cancers (Fig. S7). The toxicity enrichment analysis revealed increased levels of alkaline phosphatase (ALP). Interestingly, the serum ALP level have been found to be significantly higher in OSCC patients (*Acharya et al., 2017*).

Analysis of set 2: The pathway enrichment analysis by IPA showed that the top enriched pathways are neuroinflammation signalling, small cell lung cancer, MAPK signalling, and cell cycle: G2/M DNA damage checkpoints (Fig. 9) (Table S13). Top prognostics signature miRNAs are directly targeting several key cancer regulatory pathways such as p53 signalling, ATM signalling, EMT pathways etc., (Fig. 10). The KEGG pathways analysis also revealed

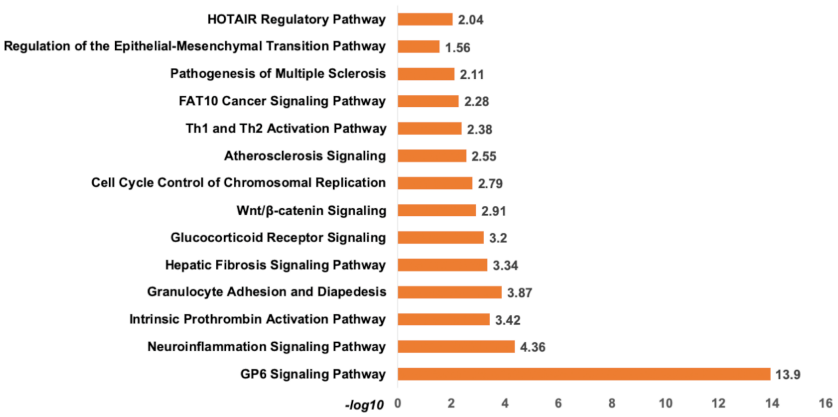

Figure 7 **The pathway enrichment analysis of signature miRNA.** The pathway enrichment analysis for the identified signatures of set 1 genes (230 mRNA). The top most enriched pathways of diagnostic signature miRNA targets genes in IPA.

Full-size DOI: 10.7717/peerj.9656/fig-7

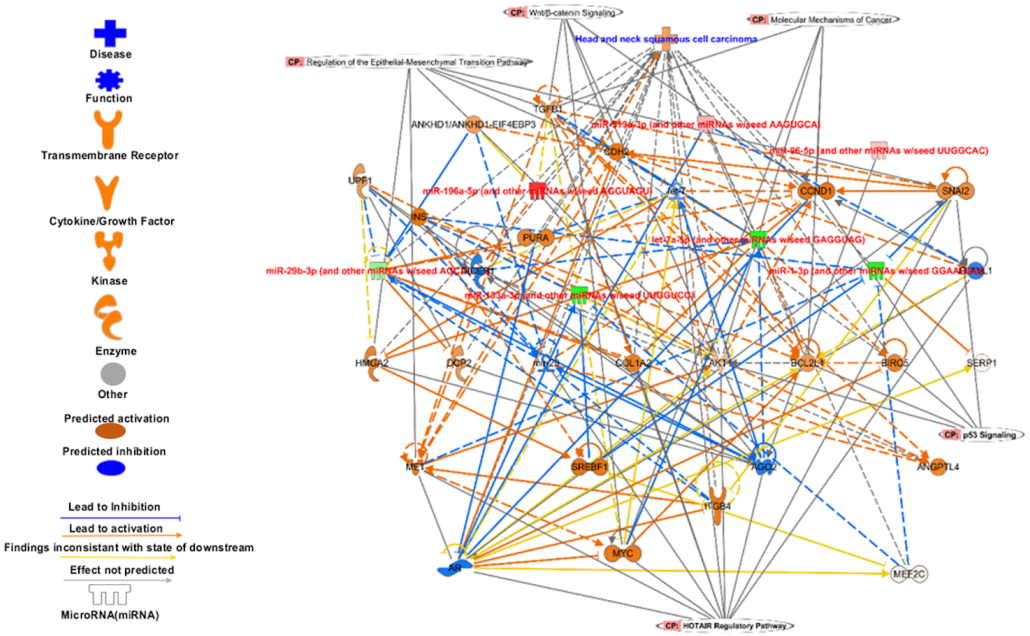

Figure 8 **The signature miRNAs targets different pathways genes.** The targets of signature miRNA in different pathways. The signature miRNA are regulating key cancer pathways e.g., Epithelial-Mesenchymal Transition, Wnt/β-catenin signalling, molecular mechanism of cancer, p53 signalling, and HOTAIR regulatory pathways.

Full-size DOI: 10.7717/peerj.9656/fig-8

that majority of the genes were associated with cancer progression such as protein digestion and absorption, ECM-receptor, transcriptional misregulation in cancer, and focal adhesion. In case of WikiPathways, the ECM and membrane receptors (ID: WP291), miR-509-3p alteration of YAP1/ECM axis (WP3967), and focal adhesion-PI3K-Akt-mTOR-signalling pathway (WP3932) were found enriched (Table S14).

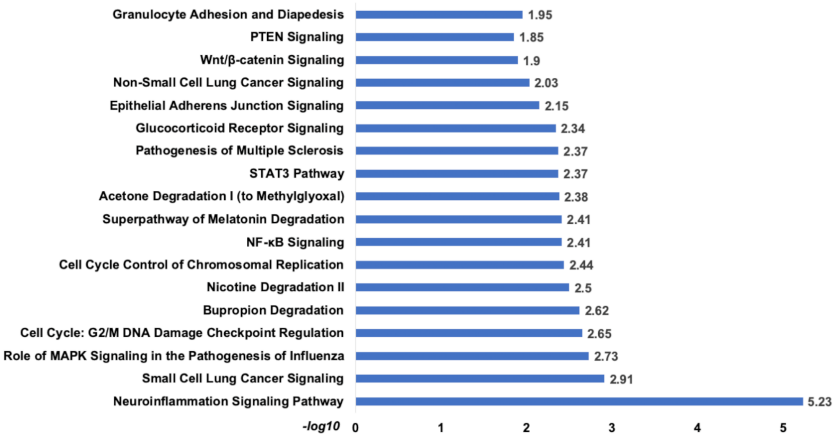

Figure 9 **The pathways enrichment analysis for the identified signatures of set 2 genes as prognostic marker.** The enriched biological pathways by IPA. The prognostics signature miRNAs are mostly enriched cancer-related pathways such as small cell lung cancer, Cell-Cycle:G2/M DNA damage checkpoint regulation, NF-kappa B signalling, etc.

Full-size DOI: 10.7717/peerj.9656/fig-9

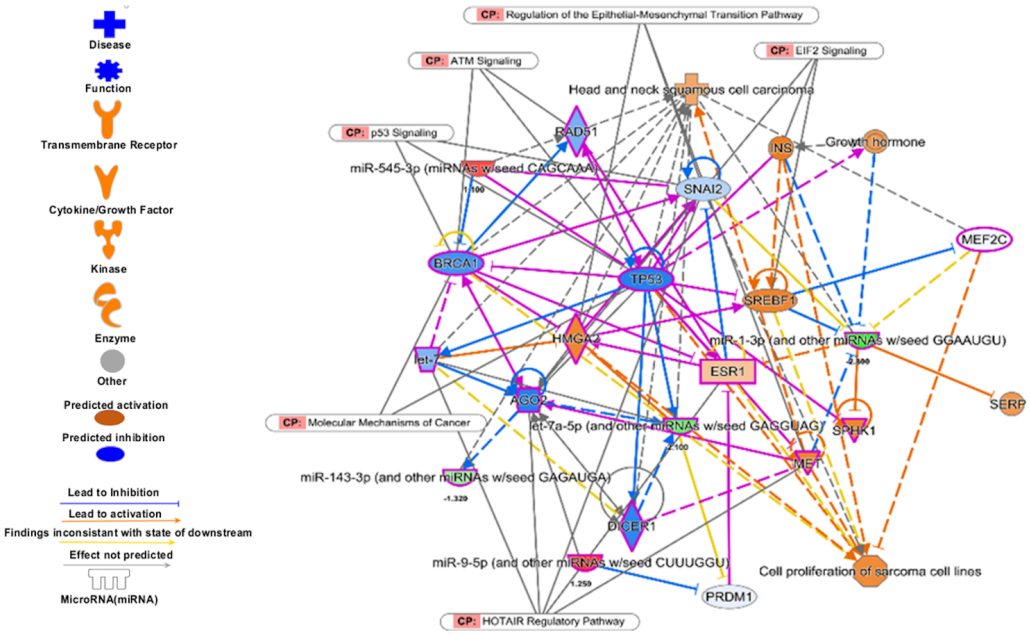

Figure 10 **The prognostics signature target different carcinogenic gene and pathways.** The molecular target of these eight signature miRNA were predominantly associated with tumor-suppressors and oncogenes in different pathways. Most of the miRNA targets are directly associated with HNSCC progression and proliferation. These miRNAs also targeting different cancer regulatory pathways i.e., p53 signalling, ATM signalling, regulation of the EMT pathways, EIF2 signalling, HOTAIR Regulatory pathways, and molecular mechanisms of cancer. The pink marked mRNAs are known to be associated with head and neck cancer.

Full-size DOI: 10.7717/peerj.9656/fig-10

The top regulatory effect networks were MITF, CD3, YAP1, let7, TGF β1, and TNF. MITF is mitochondrial transcription factor genes and have essential roles in cell differentiation, proliferation and survival of cell (*Bertolotto et al., 1998*). It is also important co-activator of TGF β (*Nijman et al., 2006*). CD3 is cluster of differentiation 3 activates in early oral premalignant (*Öhman et al., 2015*). The Hippo-YAP1 pathway is oncogenic in head and neck cancer (*Santos-de Frutos, Segrelles & Lorz, 2019*). TGF β1 dysregulation is very common in several cancers including HNSCC. It is key regulator of epithelial cell proliferation, growth factor and angiogenesis (*Rothenberg & Ellisen, 2012*). The dysregulation of TNF genes are involved in almost all types of cancer. The TNF plays an important role in cytokine production and have critical role in immune regulation (*Alsahafi et al., 2019*).

The top diseases and disorders found to be associated with the genes of this set were cancer, organismal injury & abnormalities, and hematological disorders. Additionally, top associated network functions were associated with cell death and survival, cellular development, cellular growth, and proliferation (Table S15). The miRNA-mRNA interaction analysis showed that there are 16 mRNAs which are part of cosmic consensus cancer gene panel and there are 20 transcription factors among the targets. Six mRNAs were common to transcription factors and cancer gene panels (Fig. S8). Overall enrichment analysis revealed that these miRNAs were directly or indirectly associated with the tumorigenesis and poor prognosis.

Gene ontology enrichment analysis

Set 1: The most enriched biological processes were associated with extracellular matrix proteins organization (GO:0030198), regulation of natural kill cell chemotaxis (GO:2000501) and response to alcohol (GO:0097305). While top molecular functions enriched genes were mRNA 3′-UTR binding (GO:0003730), chemokine activity (GO:0008009) and transforming growth factor beta-activated receptor activity (GO:0005024). The cellular component such as messenger ribonucleoprotein complex (GO:1990124), serine/threonine protein kinase complex (GO:1902554), gamma-tubulin large complex (GO:0000931) were found majority of genes involved in these process (Fig. S9).

Set 2: Many of the enriched GO processes are similar to set 1. However, there are notable differences as well. The most enriched genes were involved in biological process such as collagen fibril organization (GO:0030199), negative regulation of myeloid cell differentiation (GO:0045638), and negative regulation of cell differentiation (GO:0045596). In the molecular functions enriched genes were platelet-derived growth factor binding (GO:0048407), transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding (GO:0001228), cell proliferation (GO:0008283), metabolic process (GO:0008152), biological adhesion (GO:0022610). The endoplasmic reticulum lumen (GO:0005788), integral component of plasma membrane (GO:0005887), and intermediate filament (GO:0005882) were most enriched cellular component terms (Fig. S10).

DISCUSSION

The HNSCC is associated with high mortality and morbidity. It is highly heterogeneous and identification of robust, potential and reproducible biomarker is a major challenge. The TCGA provides detailed transcriptomic data with clinical information that can be gainfully used for better risk prediction and disease management. The ML techniques, riding the recent advancement in software, hardware, and sequencing technologies, are being used to identify patterns or factors (mutation/variation) for personalized diagnostics, prognostics, and therapeutics (*Kann et al., 2019*; *Topol, 2019*). The alteration in miRNA expression in tissue as well as in biological fluids has been reported to be an important biomarker in various cancers. Interestingly, during the course of this work, we came across the announcement by Toshiba company (https://www.toshiba.co.jp/rdc/rd/detail_e/e1911_06.html) about a device that can predict the presence of 13 different cancers using the miRNA expression in blood alone with high (>99%) accuracy. Thus proving beyond doubt that miRNA expression data can be utilized for the development of a simple yet highly accurate screening tool for various cancers.

The etiology and aggressive behavior of HNSCC are not well understood at the molecular level. It calls for the studies to improve the molecular understanding leading to overall better clinical outcomes. Therefore, the main purpose of this study was to identify a set of signature miRNA using ML techniques to correctly classify the early and late-stage along with other prognostic biomarkers. The main findings of the study are outlined below:

The machine learning techniques identified biomarkers with good accuracy: The analysis resulted in the identification of 6 miRNA signature which predicted the early and late stages with good accuracy on external data set. The good accuracy in classification of the samples in the external set further validated the good classification ability of the generated models. It is pertinent to note that the stage is also a significant prognostic indicator.

The evaluation of the identified prognostic signature indicated that these miRNAs are significantly associated with poor survival of the patient and thus it can be a good prognostics biomarker. The IPA, KEGG and WiKipathways analysis also revealed that target of these miRNAs are highly associated with the cancer initiation and progression. The molecular functional analysis of the identified miRNA also indicated that they directly or indirectly regulate several oncogenes, tumor suppressor genes and transcription factors.

The data balancing by enhanced sampling can improve accuracy: The small cohort and imbalanced data highlight the major challenges in this study. The SMOTE algorithm was therefore utilized to generate a balanced dataset for classification of stages (*Vafaee et al., 2018*). It was interesting to note that the models generated with oversampling of the data showed better accuracy as compared to the undersampled data. Perhaps, the undersampling of the data might have resulted in loss of important information and thus reduced accuracy.

The identified miRNAs are detectable in biological fluids: A detailed literature survey suggested that the identified miRNAs are related to diagnostics and prognostics of many cancers. Interestingly, they are also found in biological fluids (plasma, serum, etc.) in several cancers including HNSCC. The miR-29 and let-7 family miRNAs were significantly

downregulated ($p < 0.001$) in the serum of patients with high-risk oral lesions. The miR-196a was found up-regulated in the plasma of HNSCC patients. The miR-486 was found downregulated in blood samples of non-small lung cancer patients. The reduced expression of circulating miRNA-133b was also associated with the clinical stage, metastasis, and survival of breast cancer patients. The upregulation of miR-96 is associated with various cancers such as breast (*Zhang et al., 2017*), hepatic (*Musaddaq et al., 2019*), and colorectal (*Brunet Vega et al., 2013*). The deregulation of miR-519 is linked with the prognosis and diagnosis of lung cancer and it's found to be in circulating in plasma (*Wang et al., 2019*). The overall literature survey suggested that the identified miRNAs were found dysregulated in biological fluids of patients of various cancers. They can also be useful for the detection and prognosis of HNSCC in biological fluids as well, though experimental validation would be needed to prove this assumption.

CONCLUSION

In this study we have tried to identify miRNA expression patterns for classification of TNM stage (early and late stage) in HNSCC patients using machine learning algorithms (RF, svmR, adaBoost, avNNet and GBM). The dissection of miRNA expression with the help of machine learning tools provided us with miRNA signatures that can distinguish between early and late-stage tumor samples and risk to the patient with good accuracy. The five-year overall survival analysis revealed that the dysregulation of the identified miRNAs is significantly correlated with poor prognosis. We have further identified the functional roles of the identified miRNAs, their mRNA targets, pathways and process that may get perturbed in the HNSCC carcinogenesis and progression. Very interestingly, the identified miRNAs were also found to be differentially expressed in the biological fluids of patients of HNSCC patients. This opens a possibility that these miRNAs can also be proposed as non-invasive biomarkers for HNSCC, although it will need a large number of patient samples and experimental validation. We hope that these studies will help in the development of potential biomarkers and miRNA based therapies in cancer.

ACKNOWLEDGEMENTS

The authors are thankful to Director, Institute of Life Sciences, Bhubaneswar, India and Dept. of Biotechnology (DBT), Govt. of India for providing necessary facilities. The authors are grateful to all patients who provided samples to the TCGA project. The authors are thankful to Shaheerah and Pratima for critical reading of the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Sugandh Kumar received a fellowship from the University Grant Commission (UGC), Government of India. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
University Grant Commission (UGC), Government of India.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Sugandh Kumar conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Srinivas Patnaik conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Anshuman Dixit conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Raw data are available at TCGA under project ID TCGA-HNSC.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.9656#supplemental-information.

REFERENCES

Acharya S, Kale J, Rai P, Anehosur V, Hallikeri K. 2017. Serum alkaline phosphatase in oral squamous cell carcinoma and its association with clinicopathological characteristics. *South Asian Journal of Cancer* **6**:125–128 DOI 10.4103/2278-330X.214574.

Akhtar A, Hussain I, Talha M, Shakeel M, Faisal M, Ameen M, Hussain T. 2016. Prevalence and diagnostic of head and neck cancer in Pakistan. *Pakistan Journal of Pharmaceutical Sciences* **29**:1839–1846.

Alabi RO, Elmusrati M, Sawazaki-Calone I, Kowalski LP, Haglund C, Coletta RD, Makitie AA, Salo T, Leivo I, Almangush A. 2019. Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool. *Virchows Archiv* **475**:489–497 DOI 10.1007/s00428-019-02642-5.

Alexopoulos EC. 2010. Introduction to multivariate regression analysis. *Hippokratia* **14**:23–28.

Alsahafi E, Begg K, Amelio I, Raulf N, Lucarelli P, Sauter T, Tavassoli M. 2019. Clinical update on head and neck cancer: molecular biology and ongoing challenges. *Cell Death & Disease* **10**:1–17 DOI 10.1038/s41419-018-1236-z.

Avissar M, Christensen BC, Kelsey KT, Marsit CJ. 2009. MicroRNA expression ratio is predictive of head and neck squamous cell carcinoma. *Clinical Cancer Research* **15**:2850–2855 DOI 10.1158/1078-0432.CCR-08-3131.

Bertolotto C, Abbe P, Hemesath TJ, Bille K, Fisher DE, Ortonne J-P, Ballotti R. 1998. Microphthalmia gene product as a signal transducer in cAMP-induced differentiation of melanocytes. *The Journal of Cell Biology* **142**:827–835 DOI 10.1083/jcb.142.3.827.

Bianchi F, Nicassio F, Marzi M, Belloni E, Dall'Olio V, Bernard L, Pelosi G, Maisonneuve P, Veronesi G, Di Fiore PP. 2011. A serum circulating miRNA diagnostic test to identify asymptomatic high-risk individuals with early stage lung cancer. *EMBO Molecular Medicine* **3**:495–503 DOI 10.1002/emmm.201100154.

Blighe K. 2018. Enhancedvolcano: publication-ready volcano plots with enhanced colouring and labeling.

Boscolo-Rizzo P, Zorzi M, Mistro ADel, Da Mosto MC, Tirelli G, Buzzoni C, Rugge M, Polesel J, Guzzinati S, Group AW. 2018. The evolution of the epidemiological landscape of head and neck cancer in Italy: is there evidence for an increase in the incidence of potentially HPV-related carcinomas? *PLOS ONE* **13(2)**:e0192621 DOI 10.1371/journal.pone.0192621.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **68**:394–424.

Brunet Vega A, Pericay C, Moya I, Ferrer A, Dotor E, Pisa A, À Casalots, Serra-Aracil X, Oliva J-C, Ruiz A. 2013. microRNA expression profile in stage III colorectal cancer: circulating miR-18a and miR-29a as promising biomarkers. *Oncology Reports* **30**:320–326 DOI 10.3892/or.2013.2475.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**:321–357 DOI 10.1613/jair.953.

Collaboration GBoDC. 2019. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: a systematic analysis for the global burden of disease study. *JAMA Oncology* **5(12)**:1749–1768 DOI 10.1001/jamaoncol.2019.2996.

Deshpande AM, Wong DT. 2008. Molecular mechanisms of head and neck cancer. *Expert Review of Anticancer Therapy* **8**:799–809 DOI 10.1586/14737140.8.5.799.

Esteller M. 2011. Non-coding RNAs in human disease. *Nature Reviews Genetics* **12**:861 DOI 10.1038/nrg3074.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**:1189–1232.

Goeman JJ. 2010. L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal* **52**:70–84.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**:2847–2849 DOI 10.1093/bioinformatics/btw313.

Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. 2018. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics-Proteomics* **15**:41–51.

Huang G-B, Zhu Q-Y, Siew C-K. 2004. In: *Extreme learning machine: a new learning scheme of feedforward neural networks. 2004 IEEE international joint conference on neural networks (IEEE Cat No 04CH37541)*. IEEE, 985–990.

Iorio MV, Croce CM. 2012. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Molecular Medicine* **4**:143–159 DOI 10.1002/emmm.201100209.

Kann BH, Thompson R, Thomas Jr CR, Dicker A, Aneja S. 2019. Artificial intelligence in oncology: current applications and future directions. *Oncology* **33(2)**:46–53.

Kapoor A, Kumar A. 2019. Head-and-neck dermatofibrosarcoma protuberans: scooping out data even in dearth of evidence [Abstract 256]. *Cancer Research, Statistics, and Treatment* **2** DOI 10.4103/CRST.CRST_70_19.

Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. 2019. Deep learning-based survival prediction of oral cancer patients. *Scientific Reports* **9**:6994 DOI 10.1038/s41598-019-43372-7.

Komura D, Ishikawa S. 2018. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal* **16**:34–42 DOI 10.1016/j.csbj.2018.01.001.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* **42**:D68–D73 DOI 10.1093/nar/gkt1181.

Kuhn M. 2012. The caret package. Vienna: R Foundation for Statistical Computing. *Available at https://cranr-projectorg/package=caret*.

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**:R29 DOI 10.1186/gb-2014-15-2-r29.

Leemans CR, Snijders PJ, Brakenhoff RH. 2018. The molecular landscape of head and neck cancer. *Nature Reviews Cancer* **18**:269 DOI 10.1038/nrc.2018.11.

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* **2**:18–22.

Lusa L. 2013. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* **14**:106 DOI 10.1186/1471-2105-14-106.

MacFarlane L-A, Murphy R. 2010. MicroRNA: biogenesis, function and role in cancer. *Current Genomics* **11**:537–561 DOI 10.2174/138920210793175895.

Marur S, D'Souza G, Westra WH, Forastiere AA. 2010. HPV-associated head and neck cancer: a virus-related cancer epidemic. *The Lancet Oncology* **11**:781–789 DOI 10.1016/S1470-2045(10)70017-6.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research* **47**:D419–D426 DOI 10.1093/nar/gky1038.

Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. 2019. Machine learning and integrative analysis of biomedical big data. *Gene* **10**:87 DOI 10.3390/genes10020087.

Mostert B, Sieuwerts AM, Martens JW, Sleijfer S. 2011. Diagnostic applications of cell-free and circulating tumor cell-associated miRNAs in cancer patients. *Expert Review of Molecular Diagnostics* **11**:259–275 DOI 10.1586/erm.11.11.

Musaddaq G, Shahzad N, Ashraf MA, Arshad MI. 2019. Circulating liver-specific microRNAs as noninvasive diagnostic biomarkers of hepatic diseases in human. *Biomarkers* **24**:103–109 DOI 10.1080/1354750X.2018.1528631.

Network CGA. 2015. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**:576–582 DOI 10.1038/nature14129.

Nijman SM, Hijmans EM, El Messaoudi S, Van Dongen MM, Sardet C, Bernards R. 2006. A functional genetic screen identifies TFE3 as a gene that confers resistance to the anti-proliferative effects of the retinoblastoma protein and transforming growth factor- β. *Journal of Biological Chemistry* **281**:21582–21587 DOI 10.1074/jbc.M602312200.

Obermeyer Z, Emanuel EJ. 2016. Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal of Medicine* **375**:1216–1219 DOI 10.1056/NEJMp1606181.

Öhman J, Mowjood R, Larsson L, Kovacs A, Magnusson B, Kjeller G, Jontell M, Hasseus B. 2015. Presence of CD3-positive T-cells in oral premalignant leukoplakia indicates prevention of cancer transformation. *Anticancer Research* **35**:311–317.

Poddar A, Aranha R, Royam MM, Gothandam KM, Nachimuthu R, Jayaraj R. 2019. Incidence, prevalence, and mortality associated with head and neck cancer in India: protocol for a systematic review. *Indian Journal of Cancer* **56**:101–106 DOI 10.4103/ijc.IJC_416_18.

Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F, Khalili D. 2016. The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. *Medical Decision Making* **36**:137–144 DOI 10.1177/0272989X14560647.

Rothenberg SM, Ellisen LW. 2012. The molecular pathogenesis of head and neck squamous cell carcinoma. *The Journal of Clinical Investigation* **122**:1951–1957 DOI 10.1172/JCI59889.

Santos-de Frutos K, Segrelles C, Lorz C. 2019. Hippo pathway and YAP Signaling alterations in squamous cancer of the head and neck. *Journal of Clinical Medicine* **8**:2131 DOI 10.3390/jcm8122131.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**:2498–2504 DOI 10.1101/gr.1239303.

Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. 2018. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**:696–705 DOI 10.1038/s41568-018-0060-1.

Stenson KM, Brockstein BE, Ross ME. 2016. Epidemiology and risk factors for head and neck cancer. UpToDate. *Available at https://www.uptodate.com/contents/*

epidemiology-and-risk-factors-for-head-and-neck-cancer#:~:text=Head%20and%20neck%20cancer%20is,infection%20(for%20nasopharyngeal%20cancer).

Topol EJ. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **25**:44–56 DOI 10.1038/s41591-018-0300-7.

Vafaee F, Diakos C, Kirschner MB, Reid G, Michael MZ, Horvath LG, Alinejad-Rokny H, Cheng ZJ, Kuncic Z, Clarke S. 2018. A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. *NPJ Systems Biology and Applications* **4**:1–12 DOI 10.1038/s41540-017-0037-9.

Vigneswaran N, Williams MD. 2014. Epidemiologic trends in head and neck cancer and aids in diagnosis. *Oral and Maxillofacial Surgery Clinics of North America* **26**:123–141 DOI 10.1016/j.coms.2014.01.001.

Wang L, Mo H, Jiang Y, Wang Y, Sun L, Yao B, Chen T, Liu R, Li Q, Liu Q. 2019. MicroRNA-519c-3p promotes tumor growth and metastasis of hepatocellular carcinoma by targeting BTG3. *Biomedicine & Pharmacotherapy* **118**:109267 DOI 10.1016/j.biopha.2019.109267.

Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, Sander C, Cherniack AD, Mina M, Ciriello G. 2018. Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas. *Cell Reports* **23**:172–180. e173 DOI 10.1016/j.celrep.2018.03.046.

Wiegand S, Zimmermann A, Wilhelm T, Werner JA. 2015. Survival after distant metastasis in head and neck cancer. *Anticancer Research* **35**:5499–5502.

Zemel RS, Pitassi T. 2001. A gradient-based boosting algorithm for regression problems. In: *Advances in Neural Information Processing Systems*. 696–702.

Zhang K, Wang Y-W, Wang Y-Y, Song Y, Zhu J, Si P-C, Ma R. 2017. Identification of microRNA biomarkers in the blood of breast cancer patients based on microRNA profiling. *Gene* **619**:10–20 DOI 10.1016/j.gene.2017.03.038.