Artificial Intelligence based models for Predicting Head and Neck Cancer from Genomics data of Single Cells

A Project Report

Submitted by

AKANKSHA JARWAL

In partial fulfilment of the requirements

for the award of the degree of

MASTER OF TECHNOLOGY

COMPUTATIONAL BIOLOGY

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

Feb, 2023

THESIS CERTIFICATE

This is to certify that the thesis titled "**Artificial Intelligence based models for Predicting Head and Neck Cancer from Genomics data of Single Cells**" submitted by **Akanksha Jarwal**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of Masterof Technology, is a bona fide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree.

Advisor's Name
Prof. Gajendra Pal Singh Raghava
Department Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: 11th February, 2023

ACKNOWLEDGEMENTS

I would like to sincerely thank and acknowledge my M.Tech thesis supervisor **Prof. Gajendra P. S. Raghava** from Indraprastha Institute of Information and Technology, Delhi, for allowing me to conduct research under his guidance, from introducing me to the topic to helping me mature as a researcher. His patience and diligence motivated me to persevere during my research. Besides my supervisor, I would like to express my sincere gratitude to the Ph.D. scholars Anjali Dhall and Sumeet Patiyal for their constant guidance, and motivation and for playing an instrumental role in the entire research process. Lastly, I would like to thank my family and friends for their constant encouragement and support, which enabled me to pursue research.

Akanksha Jarwal
M.Tech Computational Biology

ABSTRACT

Head and Neck Squamous Cell Carcinoma (HNSC) or Head and Neck Cancer is the sixth most highly prevalent cancer type worldwide. Early detection of HNSC is one of the important challenges in managing the treatment of cancer patients. Existing techniques for detecting HNSC are costly, expansive, and invasive in nature. In this study, an attempt has been made to develop classification models using machine and deep learning techniques to discriminate HNSC and normal samples. In addition. models have been developed to predict HPV associated HNSC samples. All models in this study have been developed on two datasets (GSE181919 and GSE139324) of single-cell genomics obtained from RNA-seq technology. These models were trained on the training dataset and validated on internal and external datasets. Our deep learning models outperform machine learning models in the prediction of HNSC samples on bother datasets. We further classified these HNSC samples into HPV associated and HPV non-associated HNSC samples with high precision. In summary, this is a pilot study to understand the application of single-cell genomics in predicting HNSC and its type. In order to facilitate the scientific community a software package has been developed in Python which is available from the following URL https://webs.iiitd.edu.in/raghava/hnscpred/.

KEYWORDS: Head and Neck Squamous Cell Carcinoma; Machine Learning; Disease Diagnostics; Genetic Biomarkers; HPV

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

IIITD	Indraprastha Institute of Information Technology Delhi
HNC	Head Neck Cancer
HNSCC	Head and Neck Squamous Cell Carcinoma
HPV+	Human Papillomavirus Positive
HPV-	Human Papillomavirus Negative
GO	Gene Ontology
AUC	Area Under Curve
DT	Decision Tree
RF	Random Forest
ANN	Artificial Neural Network
IFS	Incremental Feature Selection
mRMR	Minimal Redundancy Maximal Relevance
sc RNA	Single Cell Ribosomal Nucleic Acid
MCC	Matthews Correlation Coefficient

CHAPTER 1

INTRODUCTION

Head and neck cancer, which encompasses a range of malignancies affecting the respiratory tract and upper digestive tract, is the sixth most common kind of cancer in the world. Squamous cell carcinoma is the most typical kind, however, there are many others [1]. Squamous cell carcinomas often develop in the salivary glands, larynx, oral cavity, throat, and Sino nasal tract epithelium. Even after meticulous and selective therapy, the odds of survival are decreased by the fact that most of the cases of head and neck cancer are detected at advanced (late) stages [2].

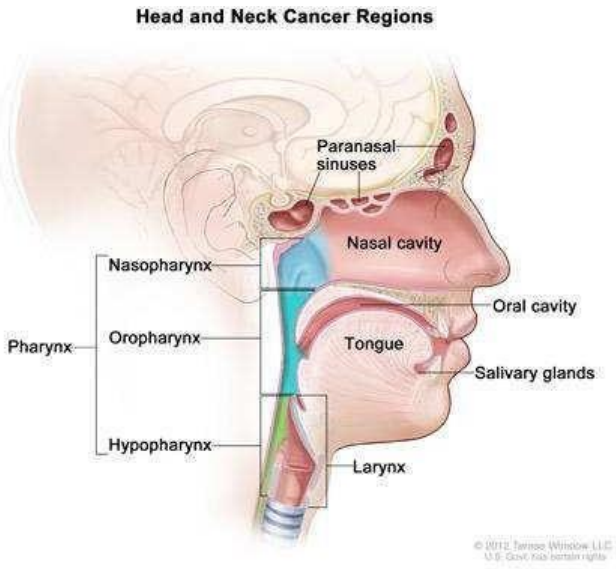

Figure 1.1: Head and neck cancer regions. Illustrates the location of paranasal sinuses, nasal cavity, oral cavity, tongue, salivary glands, larynx, and pharynx (including the nasopharynx, oropharynx, and hypopharynx. Credit: © Terese Winslow

More and more often, human papillomavirus (HPV) infection, notably HPV-16, is linked to several head and neck malignancies. Depending on the country or area, different carcinogens from smoking or excessive alcohol use have been related to the development of HNSCC. As a result, HNSCC can be classified either HPV-negative or HPV-positive [3].

Careful physical examination continues to be the critical method for early detection

because no screening tool has been demonstrated to be helpful. Worldwide, 562,328 people were diagnosed with head and neck cancer (HNC) in 2020, with a total count of 277,587 deaths due to the disease [4]. However, due to imprecise histories and indistinctive diagnostic characteristics, early identification of some HNC can be challenging. In addition to clinical and radiological investigations, the traditional diagnosis of HNC is based on histological analysis of tissue sections obtained from biopsies or surgical resections. These procedures can take a lot of time and are susceptible to mistakes in observation or interpretation, which can lead to discrepancies in cancer grading and prognostication [2]. The median age of diagnosis for non-virally associated HNSCC is 66 years, whereas the median age for HPV-associated oropharyngeal cancer and Ep- stein–Barr virus (EBV)-associated nasopharyngeal cancer is 53 years and 50 years, respectively [3].

Secondary prevention is greatly aided by biomarkers. The FDA has approved 28 biomarkers for clinical usage based on rigorous in vitro studies. However, the FDA has not approved any protein, or mutation marker for use in HNSCC diagnosis or prognosis [5]. Single-cell technology makes it possible to detect molecular alterations in individual cancer cells. This can increase the research of more specialized biomarkers with excellent resolution, leading to the development of a complete landscape of distinct cell types within tumors [6].

This study aims to predict HNSCC patients using essential biomarkers from single-cell sequencing data of multiple patients affected by Head and neck cancer. These 100 genes, identified as biomarkers in our study, can be crucial in diagnosing head and neck cancer.

1.1 Single cell Genomics

Single-cell sequencing analyses the sequencing data collected from individual cells using next-generation sequencing methods that have been refined, giving a better knowledge of the activity of a single cell in relation to its microenvironment and a higher resolution of cellular distinctions [7]. Cell-to-cell variation can be revealed by single-cell RNA sequencing or epigenetic alterations, which may aid populations in quickly adapting to new circumstances [8]. The significance of gene mosaicism as well as intra-tumor genetic heterogeneity in the genesis of cancer or response to therapy can

be uncovered by single-cell precision [9].

Reverse transcription (RT), amplification, library building, and sequencing are all processes in the existing scRNA-seq techniques, including isolating individual cells and their RNA. The reverse transcription procedure, which turns RNAs into cDNAs, occurs when individual cells are enclosed in droplets in a microfluidic device, as opposed to the earlier approaches, which split individual cells into separate wells. A DNA "bar-code" that uniquely identifies the cDNAs generated from a single cell is carried by each droplet. The cDNAs from various cells can be combined for sequencing when reverse transcription is finished since each transcript from a particular cell can be distinguished by its distinct barcode [10, 11].

The limitations of bulk-RNA sequencing were completely solved by single-cell RNA sequencing (scRNA-seq), which also made it possible to determine the distribution within each gene's expression profiles all across the entire cell population. It can now respond to basic biological queries such as what kind of cells are in the tissue, what activities these cells perform, and how these processes differ from those of healthy tissues. Understanding cell-type-specific information can help in the identification of novel or uncommon cell types, understanding cell differentiation over development, and determining cell composition in healthy and sick tissues.

1.1.1 Challenges with scRNA-seq data

ScRNA-seq data offers considerable concerns due to the low initial material per cell. As a consequence, it contains a lot of zeros, and the data is generally sparse. It's feasible that the zeroes in the dataset are mistaken. Genes that are expressed in cells but are unable to detect because of technological limitations are termed "dropouts" or "fake" zeros. Genes not expressed in cells are referred to as "real" zeros. This results in unintended diversity in cells that were not brought on by biological variation but rather by technological issues, such as the gene not being sufficiently amplified by PCR. Normalization, meanwhile, can help to alleviate this problem. Batch effects are yet another concern with single-cell RNA sequencing data. Data integration or normalization is

among the most crucial steps in single-cell data analysis. As several datasets from diverse laboratories are assembled and sequenced using numerous methods and tools to result in a single sizable reference dataset, an effect can be described as the batch effect. In light of the procedural noise this introduces into the data, it becomes increasingly challenging to determine the biological heterogeneity within [12, 13].

CHAPTER 2

PREDICTION

2.1 Workflow

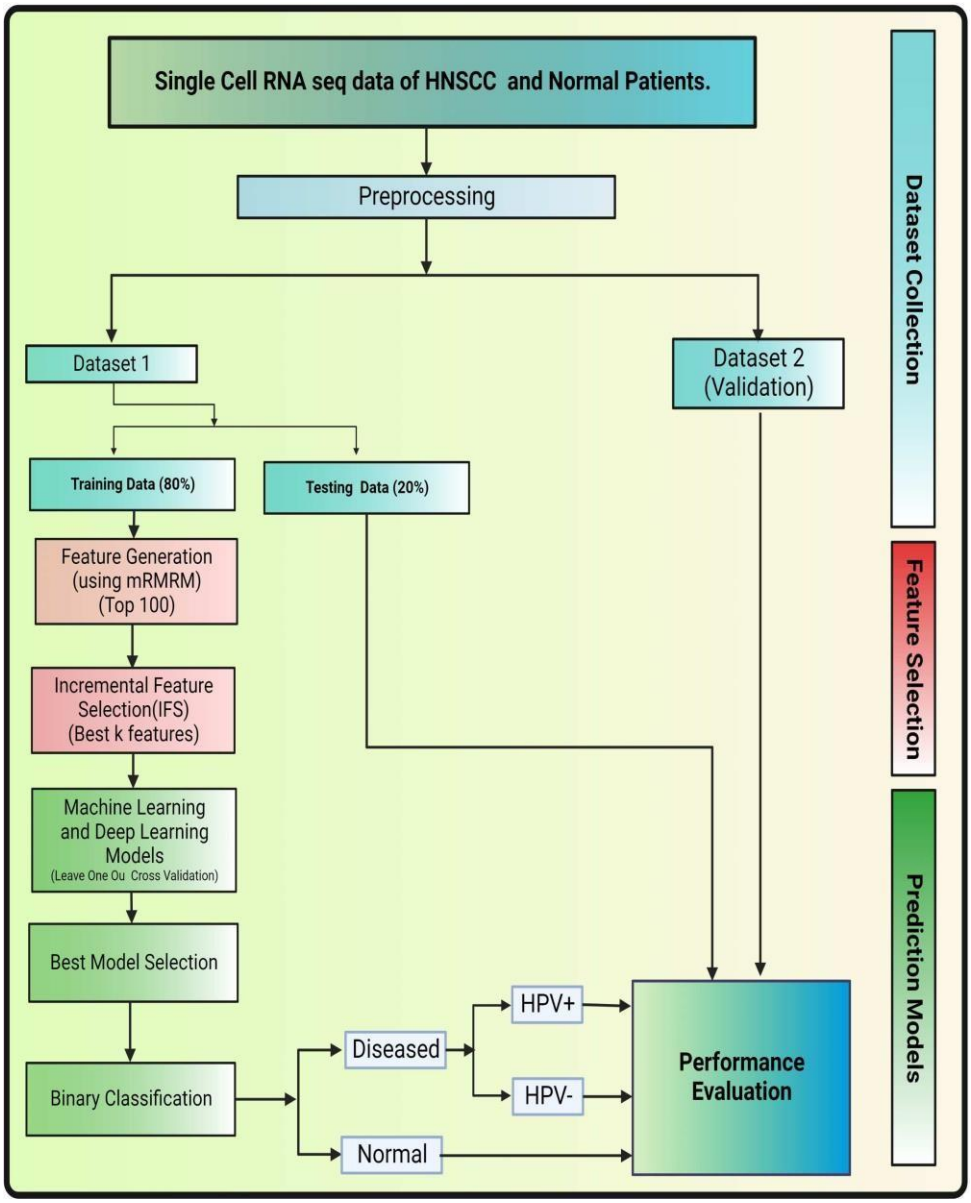

Figure 2.1: Workflow of data collection to prediction

2.2 Material and Methods

2.2.1 Dataset Preparation

a) Data Collection

Two databases with the GEO accession numbers GSE139324 [14] and GSE181919 were retrieved from NCBI GEO (Gene Expression Omnibus). Both HNSCC (diseased) and normal patients are included in these databases.

b) Data description

The dataset consists of single-nucleus RNA sequencing of 131,224 single cells from peripheral and intratumoral CD45+ immune populations from HPV– and HPV+ HNSCC and healthy donors.26 peripheral blood mononuclear cells (PBMC), 26 tumor-infiltrating immune cells from HNSCC patients, 6 peripheral blood mononuclear cells from healthy donors, and 5 tissue-resident immune cells from healthy donor tonsils were part of the 63 samples analyzed by scRNAseq in the first dataset (GSE139324) [14]. The sequencing model used for sequencing was Illumina NextSeq 500, and further data processing, such as DE multiplexing, has been done using Cell Ranger mkfastq.

In the second dataset, single-cell transcriptome profiling was performed on 37 tissue samples, including cancer (n = 28) and non-tumor surrounding normal tissue (n = 9). Illumina HiSeq 4000 was used as an instrumental model. CellRanger 2.1.1 (10x Genomics) was used to analyze and match the sequencing of this data to the human reference genome.

c) Data Pre-Processing

To begin with, the entire processing of data and analysis were carried out in a Jupyter notebook using the Python 3 environment. Since the data given by Cell Ranger Pipeline was in sparse format, it consisted of only non-zero items to reduce file size. Using the os, CSV, gzip, and scipy.io modules, we turned the data into a matrix in the form of a data frame.

- Data loading and filtering
  The information from each patient was likewise transformed into a data frame, and the dataset was pre-processed to eliminate unnecessary columns and cells. The cells having all 0s were filtered, and the genes that had no mapped expression readings to more than 80% of the cells were eliminated. Following that, labels of 0s and 1s were placed on the filtered data frames. Individuals with HN-SCC disease were labeled with 0's, whereas healthy patients were labeled with 1's. Similar steps have been performed with further classified categories of HPV,ie, HPV+ and HPV-.

- Normalization
  The sequencing depth affects the range of values for the features, which necessitates normalizing the count data before doing any sort of analysis. We used scanpy.pp.normalize_total package to conduct CPM (counts per million) normalization on this data and then performed a log transformation to give the values ofhighly expressed and weakly expressed genes equal weight.

- Data Partition
  The dataset was firstly divided into two sections for validation and training into 80%-20% ratio. The entries from both the classes of HNSCC and Normal were eventually amalgamated from all the training data frames to create a sizable data frame. Similar steps have been performed with further classified categories of HPV, i.e., HPV+ and HPV-.

2.2.2 Feature Selection

The differentially expressed genes (DEGs) in cells could be found using a variety of statistical techniques. In such circumstances, however, the cell-to-cell interaction was neglected. Following DEG analysis, a significant amount of DEG was produced, making them suitable as biomarkers. In order to do this, we employed the mRMR gene selection algorithm, which stands for Minimum Redundancy and Maximum Relevance. It is a feature selection method that helps us choose traits that are highly correlated with the output class but not among themselves [15].

The key advantage of utilizing mRMR is that it is created to discover the smallest significant feature subset out of the entire provided features that is to seek the mini- mal subset of features that has the greatest predictive potential. However, most of the other feature selection approaches employ an all-relevant strategy that also focuses on identifying all the traits that are somewhat related to the extracted features.

Redundancy between genes is taken into account by the mRMR technique in addition to relationships amongst genes and samples. Only the most relevant gene will be considered if numerous genes are identical. For various biomedical feature selection

issues, notably in single-cell RNA-Seq analysis, this strategy has been demonstrated to be successful and is often utilized [16].

A reasonable criterion for choosing biomarkers was not produced by any other statistical approach since the single-cell sample data was so large and sparse. The optimal subset of a small number of non-redundant biomarkers may thus be chosen using the mRMR approach for single-cell data analysis. K value of 100 is used to pick the top 100 ranking features, which was discovered by mRMR from the dataset's total of 2604 expressed genes. Utilizing various Machine Learning and deep learning approaches, these 100 genes were therefore utilized to categorize the cells into diseased and normal groups.

2.2.3 Machine Learning Model

Different machine-learning models have been created to aid in categorizing the data. Many different methods have been employed in bioinformatics to address classification related challenges. Thus, we've included approaches like neural networks, decision trees, extra tree classifiers, random forests, logistic regression, and the K-neighbor classifier.

a) **Artificial Neural network**: - In this method, biological neuron networks are usedas reference [17]. Networks are composed of multiple layers, and each layer has a number of nodes (or neurons) that support decision-making. So every node starts off with a random weight that is later adjusted and set to the most appropriate value as the learning process progresses [18]. The sample's projected label (Diseased or Normal) is the output's eventual outcome.

b)**Logistic Regression**: - It is comparable to the linear regression algorithm, which transforms each sample's likelihood of prediction into a no or yes choice [19]. It makes use of the idea of the sigmoid function, which is used to calculate the likelihood that a sample belongs to a specific class [20].

c) **Decision trees**: - A decision-making technique known as a decision tree uses a tree-like depiction of alternatives, and their outcomes [21]. To decide which catego

rization to use, decision trees use the concepts of entropy, and information gain [22].

d)**Random Forest**:- A classification system made up of multiple decision trees is called the random forest approach [23]. In an effort to create a statistically independent forest of trees whose prediction by itself is more reliable than that of each individual tree, it makes use of the concepts of bagging and feature randomization [24].

e) **Extra Tree Classifiers**: This decision tree classifier uses an ensemble approach. The results of several de-correlated trees are combined [25]. We eventually reach the desired categorization outcome. Although choice trees are built differently from random forests, the underlying concept is the same.

f) **K neighbor's classifier**: - It is a feature selection method that helps us pick traits that are highly correlated with the output class but not highly correlated with one another [26]. It is an algorithm without parameters.

2.2.4 Cross Validation

The dataset was primarily composed of training data, which made up 80% of it, and independent validation data, which made up the remaining 20%. In the LOOCV (Leave One out Cross Validation) approach, the training data were further split into training and testing data, and thus the mean of the results for every fold of cross-validation was calculated. The whole training set is separated into N equivalent folds using the LOOCV technique, with (N-1) being utilized for training and the single fold being used for testing. Each fold serves as testing data for the technique's N iterations. This is a common practice in many types of study [27]. Additionally, while one dataset served as training, the other two were utilized as validation datasets for one another.

2.2.5 Model Architecture

To categorize the samples for this investigation according to their diagnoses, the special has created an artificial neural network model. Three hidden layers and an output layer make up the neural network with that. Additionally, a dropout of 0.5 is implemented at each step to lessen neural network overfitting.

Biological neuron networks served as the basis for this strategy. Artificial neurons,

which are constructed from a network of connected units or nodes and are conceptually similar to the neurons in the human brain, are used to build ANNs. They consist of several layers, and inside each layer are multiple nodes (or neurons) that supportin decision-making. The anticipated label (Diseased or Normal) of the sample is the final result.

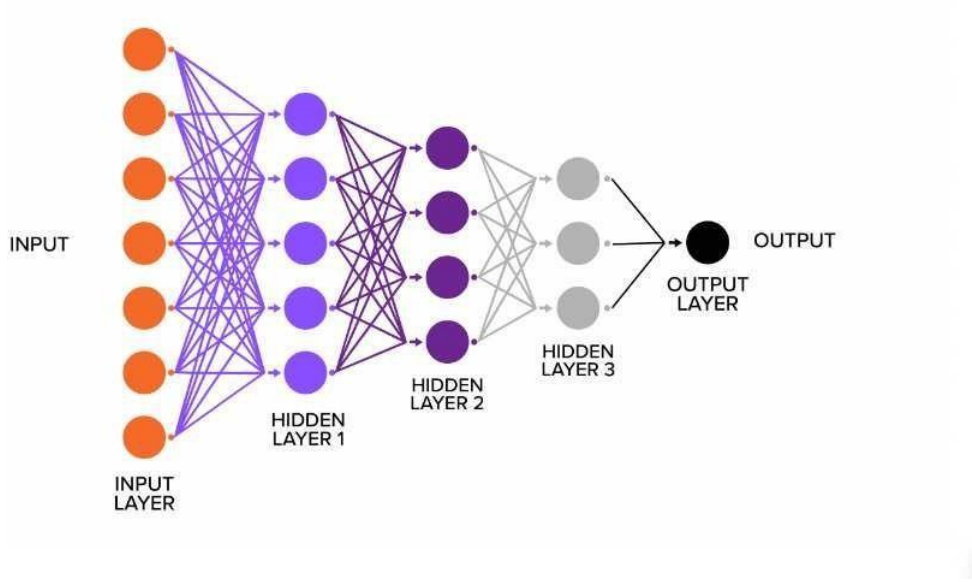

Figure 2.2: The diagram shows ANN model architecture. Credit: smartboost.com

2.2.6 Evaluation Parameter

To evaluate how good our classifier is, several metrics have been chosen to show the results. Our performance is evaluated on LOOCV. Metrics used for evaluation are:

a) **Accuracy**: It tells how many of our predictions are correctly predicted in the dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

b) **Sensitivity**: It is also called the true positive rate. It is expressed as the ratio of the number of times a sample was classified as positive when it was actually positive to the total number of positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

c) **Specificity**: The true negative rate is a different name for it. It is calculated as the ratio of the total number of negative samples to the number of times a sample was incorrectly labeled as negative.

$$Specificity = \frac{TN}{FP + TN} \tag{3}$$

d) **Precision**: It is also called the Positive Predictive value. It is expressed as the ratio of a total number of times a sample was classified as positive when it was actually positiveto the total number of times the classifier labeled a sample positive.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

e) **ROC-AUC**: By graphing the True positive rate and the False positive rate, the Receiver Operating Characteristics (Area Under Curve) visual tool helps to demonstrate the predictive ability of a classifier.

$$TruePositiveRate(TPR) = Recall = \frac{TP}{TP + FN} \tag{5}$$

$$FalsePositiveRate(FPR) = 1 - Specificity = \frac{FP}{TN + FP} \tag{6}$$

f) **F1 Score**: A classifier's precision and recall are combined into one metric by the F1-score in statistics of the classification model by calculating their harmonic means.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

(where TP, FN, FP, and TN stand for true positive, false negative, false positive, and true negative, respectively.)

2.2.7 Packages and Tools

Python 3.9.13 was used to code the whole data analysis and prediction pipeline. The ’scikit-learn’ (sklearn) [28] library was being used for machine learning, while the mRMR classification module [15] was utilized for feature selection. Multiple scanpy libraries [29] for preprocessing and filtering processes, along with the "pandas" library for loading and preprocessing data. Tensor flow and Keras were also employed as the library to generate ANN models [30].

CHAPTER 3

RESULT

3.1 Selection of Important Genes

In this study, the scRNA seq two databases with the GEO accession numbers GSE139324 [14] and GSE181919 from NCBI GEO (Gene Expression Omnibus) were analyzed, which contains both HNSCC (diseased) and normal patients. After that, pre-processing was done with the steps of filtration, normalization, and data partition. Our primary goal is to find features (genes) that were closely connected to the sample classes and were not duplicates of the other features. In order to create a robust and data-independent classification or regression model, feature selection is frequently employed to reduce the enormous number of biological features [31]. The effectiveness of the feature selection technologies has a significant impact on the construction of the illness diagnosis panel to ensure the ideal number of features to choose as biomarkers after applying mRMR to ANN classifiers. In order to determine the ideal number of genes that may serve as biomarkers, we created 100 distinct ANN classifiers based on the top 100 mRMR extracted genes and used an incremental feature selection technique. First, the top gene was used to train the model, and accuracy was then assessed. Next, a new feature was added to the existing feature set, and a new set was obtained and sent to the ANN classifier after each cycle. The accuracy of the newly constructed ANN classifiers was therefore assessed using LOOCV. In order to assess the accuracy of all the subsets, models were then trained using further subsets of the top 3, top 4,.. top 100 genes. The accuracy value acquired from all subsets was depicted on the y-axis along with the IFS curve with all combinations of gene sets (features) as illustrated in plot 3.1 for HNSCC model and 3.2 for HPV model. Both plots show that when we take all 100 genes, only we get optimal results; less than that can make us compromise on the results. These 100 genes were therefore used to classify the cells into diseased and normal categories as well as in the further classification of diseased categories in HPV+ and HPV- classes using various machine learning and deep learning methodologies.

The top 100 genes we identified using the mRMR technique are displayed in the table 3.1.

Table 3.1: Top 100 genes extracted from mRMR

SNo.	Genes	SNo.	Genes	SNo.	Genes	SNo.	Genes
01	PLAC9	26	SDPR	51	SOD3	76	PSME1
02	UBB	27	RELB	52	TNIP1	77	CILP
03	ACKR1	28	CYBA	53	LIMD2	78	CD7
04	AQP7	29	TIMP3	54	RPS15A	79	EBF1
05	FXYD1	30	RPS19	55	DUSP4	80	NDUFB11
06	BTG1	31	CREM	56	SPARCL1	81	ICAM3
07	B2M	32	NOVA1	57	SCARA5	82	MGP
08	CFD	33	RPS26	58	HLA-B	83	SLC16A3
09	LTBP4	34	PDGFRL	59	FBLN2	84	MEG3
10	RPS11	35	HLA-C	60	IDH2	85	VOPP1
11	MFAP4	36	CXCL12	61	ANGPTL1	86	TXNIP
12	ISG20	37	TGFBR3	62	HLA-A	87	RPS15
13	SARAF	38	RAC2	63	FHL1	88	SSPN
14	RPL28	39	SERP1	64	EZR	89	UBC
15	ABCA8	40	VIT	65	GPSM3	90	BMP4
16	YPEL5	41	RHOG	66	F10	91	TIGIT
17	GSN	42	ADH1B	67	EEF1A1	92	ADAR
18	IL2RG	43	RHOH	68	CYTIP	93	LRRN4CL
19	OAZ1	44	COPE	69	FBLN5	94	SUB1
20	DPT	45	DCN	70	CIB1	95	GDF10
21	RPLP1	46	CNN2	71	CXCR4	96	WIPF1
22	TNXB	47	REL	72	ATP5E	97	ABI3BP
23	CD37	48	CD34	73	FIGF	98	FAM177A1
24	ARPC3	49	OST4	74	PLPP3	99	EPSTI1
25	CYBRD1	50	PRELP	75	PIM3	100	PTGIS

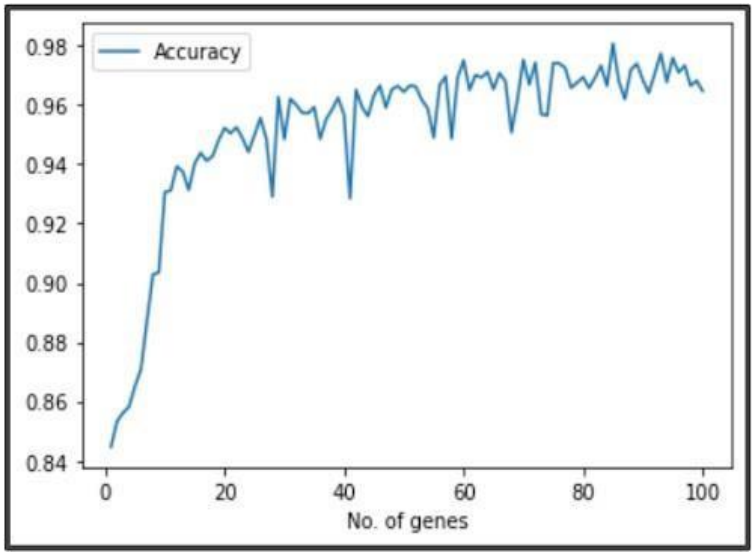

Figure 3.1: IFS accuracy graphs representing accuracy vs gene subsets.

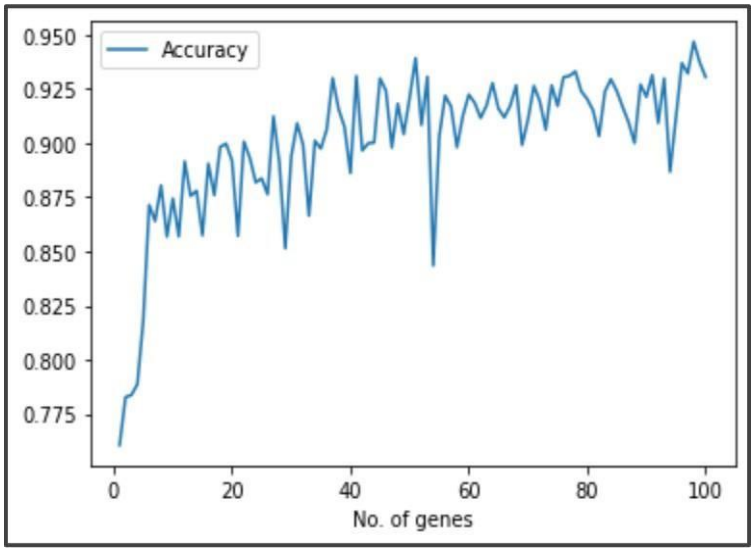

Figure 3.2: IFS accuracy graphs representing accuracy vs gene subsets.

3.2 Performance of prediction model

For HNSCC classification models

1. Training
   Many Deep and Machine Learning models were trained using the genes listed in the above table. Following training, the table 3.2 below for dataset GSE181919 and table 3.3 for dataset GSE139324" displays how well these models performed. Table 3.2 shows that the Deep Learning Model had high accuracy of 98% when compared to other machine learning models like decision trees, logistic regression, etc. In table 3.3, the Deep Learning Model has a high accuracy of 84 percent when compared with other models. As a result, when subjected to other machine learning approaches, deep learning performed noticeably better in the training part's prediction than other models.

2. Internal Validation
   The internal testing results achieved for the model trained on dataset GSE181919 in table 3.2 were 92% accuracy and 77% accuracy on testing for the model trained for dataset GSE139324. AUC-ROC achieved on the validation set was 0.91 and 0.51, respectively, on the Deep learning ANN model. Here also, it can be seen that other models failed as they were predicting with low accuracy.

3. External Validation
   In the external validation part, we do cross-testing, as we take some samples of dataset GSE139324 as testing for the model trained on dataset GSE181919 and vice versa. Therefore, a model trained for the GSE181919 dataset gives the results shown in table 3.2 on dataset GSE139324, in which again the Deep Learning modelis showing high accuracy of 69%. Vice versa, in table 3.3a model trained for the dataset GSE139324 and using samples of dataset GSE181919 as an external validation gives the accuracy of 64.89%. Hence, the Deep Learning ANN model outperforms all other models with high accuracy, whereas other models fail to predict sample labels with very low accuracy and low AUC-ROC value.

Table 3.2: Performance machine learning and deep learning models on training, internal and external validation datasets of GSE181919 dataset

HNSCC GSE181919 Dataset							
Training Evaluation Parameters							
Models	Accuracy	MCC	AUC-ROC	Sensitivity	Specificity	Precision	F1 Score
Decision Tree	0.93	0.85	0.93	0.95	0.91	0.94	0.94
Random Forest	0.96	0.92	0.96	0.98	0.94	0.96	0.97
Logistic Regression	0.92	0.84	0.92	0.95	0.88	0.92	0.94
XGBClassifier	0.92	0.83	0.92	0.92	0.92	0.95	0.93
ExtraTree Classifier	0.97	0.80	0.90	0.91	0.89	0.91	0.91
K Neighbors classifier	0.93	0.85	0.92	0.96	0.89	0.93	0.94
Deep Learning Model	0.99	0.93	0.97	0.98	0.96	0.97	0.98
Internal Validation Evaluation Parameters							
Models	Accuracy	MCC	AUC-ROC	Sensitivity	Specificity	Precision	F1 Score
Decision Tree	0.85	0.70	0.83	0.97	0.70	0.80	0.88
Random Forest	0.85	0.71	0.83	0.99	0.68	0.79	0.88
Logistic Regression	0.79	0.60	0.77	0.98	0.56	0.73	0.84
XGBClassifier	0.86	0.73	0.85	0.98	0.71	0.81	0.89
ExtraTree Classifier	0.86	0.74	0.85	0.99	0.71	0.81	0.89
K Neighbors classifier	0.83	0.68	0.81	0.98	0.65	0.77	0.87
Deep Learning Model	0.92	0.82	0.91	0.94	0.89	0.94	0.94
External Validation Evaluation Parameters							
Models	Accuracy	MCC	AUC-ROC	Sensitivity	Specificity	Precision	F1 Score
Decision Tree	0.66	-0.05	0.48	0.75	0.22	0.83	0.79
Random Forest	0.66	-0.04	0.48	0.75	0.22	0.83	0.79
Logistic Regression	0.74	0.23	0.64	0.79	0.49	0.89	0.84
XGBClassifier	0.49	-0.04	0.48	0.50	0.47	0.83	0.62
ExtraTree Classifier	0.69	0.01	0.64	0.79	0.49	0.89	0.81
K Neighbors classifier	0.64	0.03	0.52	0.70	0.34	0.85	0.76
Deep Learning Model	0.74	0.23	0.66	0.82	0.60	0.81	0.84

Moreover, we can also state that after comparing all the tables, the model trained for dataset GSE181919 is performing well in comparison with another model that was trained for GSE139324.

For HPV Classification

1. Training
   The genes mentioned in the above table were used to train multiple Machine Learning and Deep Learning models. After training, these models' performances are shown in the table below in table 3.4. Here in table 3.4, we can see Deep Learning Model was showing high accuracy of 99.1% with AUC-ROC 0.995 as

Table 3.3: Performance machine learning and deep learning models on training, internal and external validation datasets of GSE139324 dataset

HNSCC GSE139324 Dataset							
Training Evaluation Parameters							
Models	Accuracy	MCC	AUC-ROC	F1 Score	Sensitivity	Specificity	Precision
Decision Tree	0.76	0.11	0.56	0.86	0.86	0.26	0.86
Random Forest	0.81	0.10	0.54	0.90	0.95	0.12	0.85
Logistic Regression	0.82	0.19	0.58	0.90	0.94	0.22	0.86
XGBClassifier	0.78	0.08	0.54	0.88	0.90	0.17	0.85
ExtraTree Classifier	0.80	0.16	0.57	0.89	0.92	0.22	0.86
K Neighbors classifier	0.80	0.09	0.53	0.89	0.94	0.13	0.85
Deep Learning Model	0.85	0.11	0.63	0.45	0.95	0.31	0.86
Internal Validation Evaluation Parameters							
Models	Accuracy	MCC	AUC-ROC	F1 Score	Sensitivity	Specificity	Precision
Decision Tree	0.76	0.14	0.57	0.86	0.87	0.27	0.84
Random Forest	0.79	0.03	0.51	0.88	0.96	0.07	0.82
Logistic Regression	0.78	0.04	0.52	0.87	0.93	0.10	0.82
XGBClassifier	0.77	0.03	0.51	0.87	0.93	0.10	0.82
ExtraTree Classifier	0.79	0.07	0.53	0.88	0.94	0.11	0.83
K Neighbors classifier	0.79	0.12	0.55	0.88	0.94	0.15	0.83
Deep Learning Model	0.78	0.09	0.52	0.20	0.92	0.11	0.84
External Validation Evaluation Parameters							
Models	Accuracy	MCC	AUC-ROC	F1 Score	Sensitivity	Specificity	Precision
Decision Tree	0.60	0.11	0.55	0.71	0.85	0.25	0.62
Random Forest	0.59	-0.01	0.50	0.74	0.10	0.01	0.59
Logistic Regression	0.56	-0.11	0.48	0.72	0.95	0.02	0.58
XGBClassifier	0.55	-0.08	0.48	0.70	0.90	0.07	0.58
ExtraTree Classifier	0.58	-0.03	0.50	0.74	0.99	0.01	0.59
K Neighbors classifier	0.58	-0.05	0.50	0.73	0.98	0.02	0.59
Deep Learning Model	0.55	0.00	0.51	0.03	0.98	0.02	0.64

compared to other machine learning models such as logistic regression, decision trees, etc. Therefore, compared to other Machine Learning techniques Deep learning technique performed significantly well in the prediction of the training part as compared to other models.

2. Internal Validation
   The internal testing evaluation parameter results achieved for the model trained in table 3.4 were 84% accuracy. AUC-ROC achieved on the validation set was 0.83 on the Deep learning ANN model. Here also, it can be seen that other models failed as they were predicting with low accuracy.

3. External Validation
   In the external validation part, we do cross-testing, as we take some samples

Table 3.4: Performance machine learning and deep learning models on training, internal and external validation datasets of HPV dataset

HPV Dataset							
Training Evaluation Parameters							
Models	Accuracy	MCC	AUC-ROC	F1 Score	Sensitivity	Specificity	Precision
Decision Tree	0.75	0.50	0.75	0.78	0.79	0.71	0.77
Random Forest	0.82	0.63	0.81	0.84	0.87	0.75	0.81
Logistic Regression	0.84	0.67	0.84	0.86	0.86	0.82	0.85
XGBClassifier	0.77	0.52	0.76	0.79	0.81	0.71	0.77
ExtraTree Classifier	0.84	0.68	0.84	0.86	0.88	0.8	0.84
K Neighbours classsifier	0.80	0.59	0.79	0.82	0.84	0.75	0.8
Deep Learning Model	0.991	0.980	0.995	0.992	0.989	0.993	0.995
Internal Validation Evaluation Parameters							
Models	Accuracy	MCC	AUC-ROC	F1 Score	Sensitivity	Specificity	Precision
Decision Tree	0.69	0.35	0.65	0.74	0.76	0.59	0.72
Random Forest	0.84	0.88	0.80	0.83	0.91	0.92	0.94
Logistic Regression	0.80	0.54	0.75	0.85	0.9	0.61	0.81
XGBClassifier	0.82	0.61	0.81	0.86	0.87	0.74	0.86
ExtraTree Classifier	0.46	-0.11	0.76	0.52	0.58	0.33	0.47
K Neighbours classsifier	0.49	-0.06	0.55	0.55	0.58	0.38	0.52
Deep Learning Model	0.84	0.70	0.83	0.88	0.98	0.68	0.79
External Validation Evaluation Parameters							
Models	Accuracy	MCC	AUC-ROC	F1 Score	Sensitivity	Specificity	Precision
Decision Tree	0.38	-0.18	0.42	0.35	0.27	0.57	0.53
Random Forest	0.35	-0.21	0.41	0.26	0.18	0.65	0.47
Logistic Regression	0.44	-0.13	0.44	0.50	0.43	0.45	0.58
XGBClassifier	0.43	-0.03	0.49	0.38	0.28	0.71	0.62
ExtraTree Classifier	0.37	-0.16	0.43	0.29	0.2	0.66	0.51
K Neighbours classsifier	0.38	-0.15	0.44	0.33	0.23	0.64	0.53
Deep Learning Model	0.69	-0.05	0.48	0.81	0.82	0.15	0.81

of HPV dataset GSE139324 as testing for the model trained on HPV dataset GSE181919. Therefore, a model trained for the GSE181919 HPV dataset gives the results shown in table 3.4 on HPV dataset GSE139324, in which again the Deep Learning model is showing high accuracy of 69% with AUC-ROC 0.48, Sensitivity 0.82, F1 score 0.81 and MCC -0.05. Hence, the Deep Learning ANN model outperforms all other models with high accuracy, whereas other models fail to predict sample labels with very low accuracy and low AUC-ROC value.

3.3 Patient wise analysis

Since we had the data in a patient-wise format, we decided to analyze how accurately our best-selected Deep Learning Model (ANN) could predict the patient's diagnosis. Each patient's complete single-cell RNA seq profiling was given to the model to predict the percentage of 1's and 0s in it. 1's label denotes the number of diseased cells, and 0's denotes the number of normal cells. A graphical representation of the amount of predicted diseased cells and normal cells for each sample of HNSCC test and Normal Test patient is shown in Figure 3.3 and 3.4. In these bar graphs, we can see that we are getting more than 80% 1s in diseased patients and more than 70% 0s in Normal patients. Also, HPV+ and HPV- patients are shown in figure 3.6 and 3.5 where 1's percentage shows the HPV- patient and 0s percentage denotes HPV- patients. Of those, HPV- patients 1s percentage 80 and HPV- patients have 0s percentage of 60.

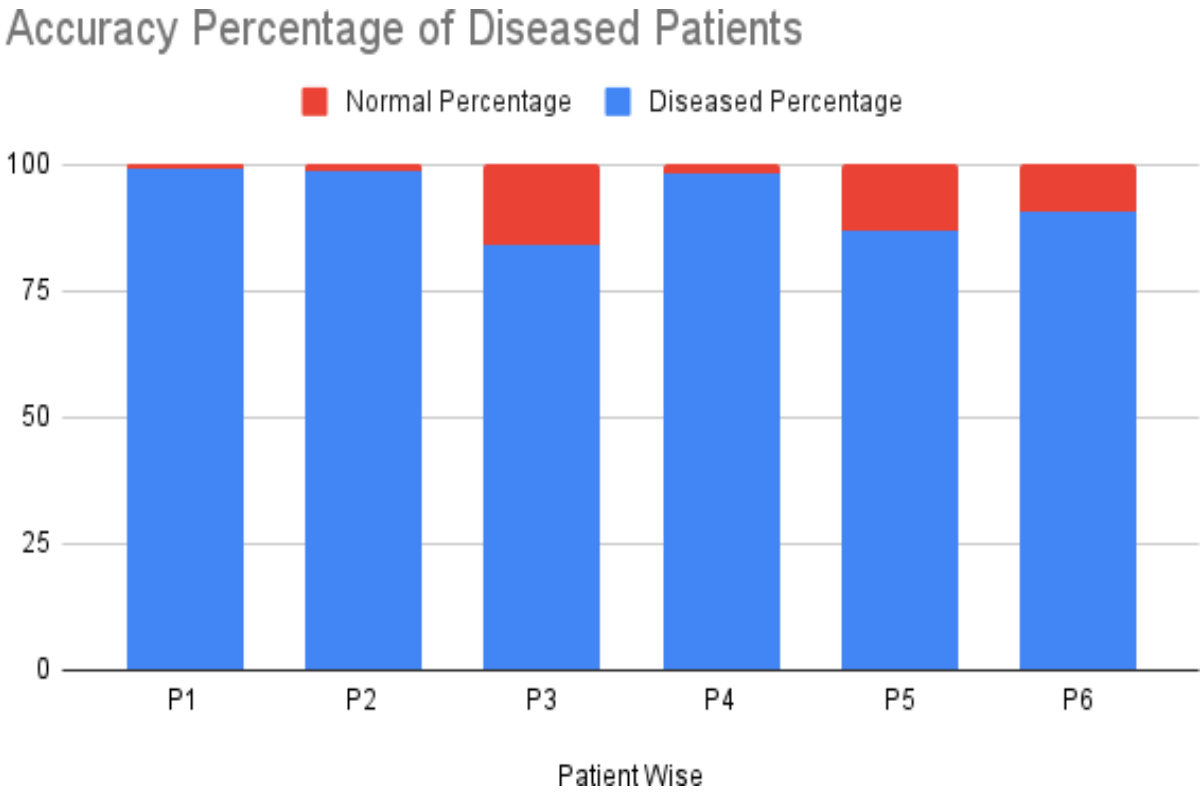

Figure 3.3: Diagrammatic representation of normal and diseased cells in HNSCC Patients

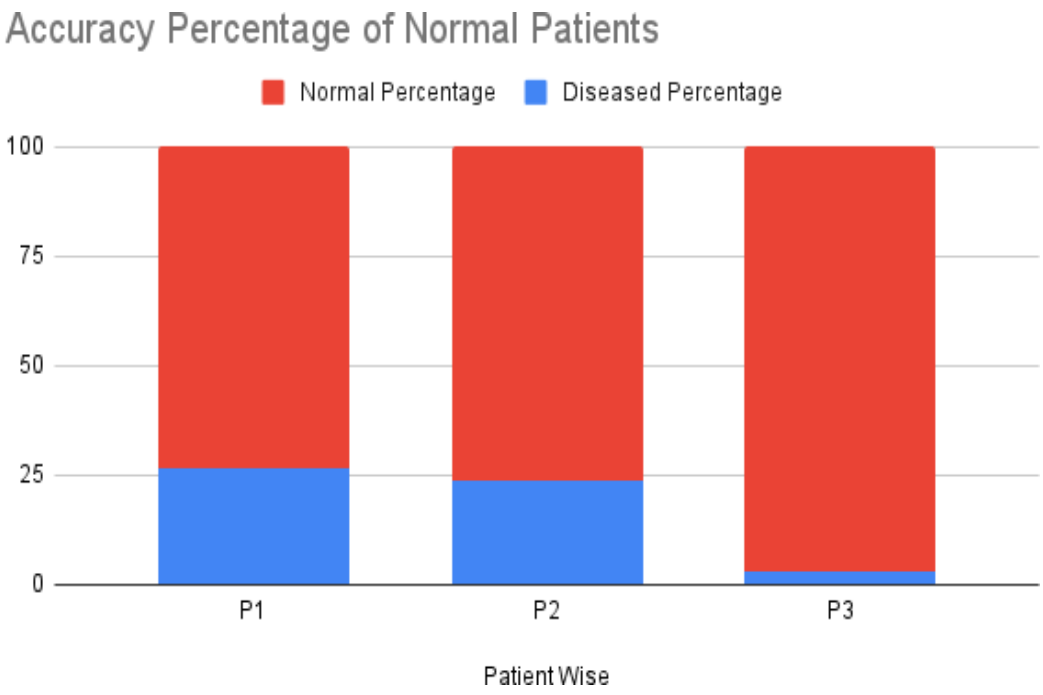

Figure 3.4: Diagrammatic representation of normal and diseased cells in Normal Patients

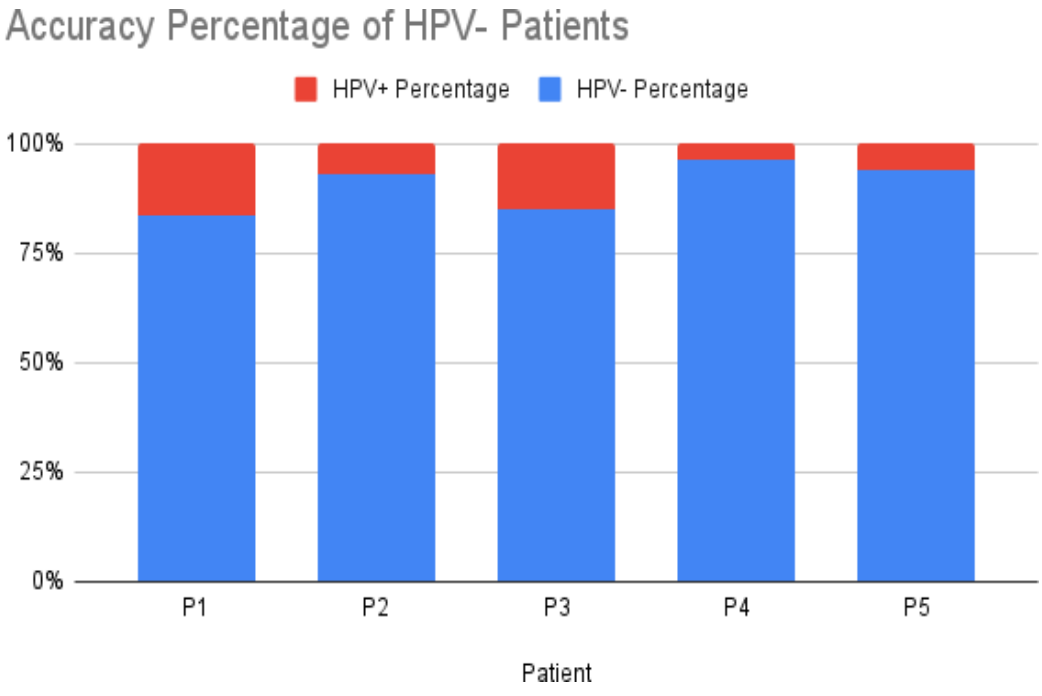

Figure 3.5: Diagrammatic representation of normal and diseased cells in HPV- Patients

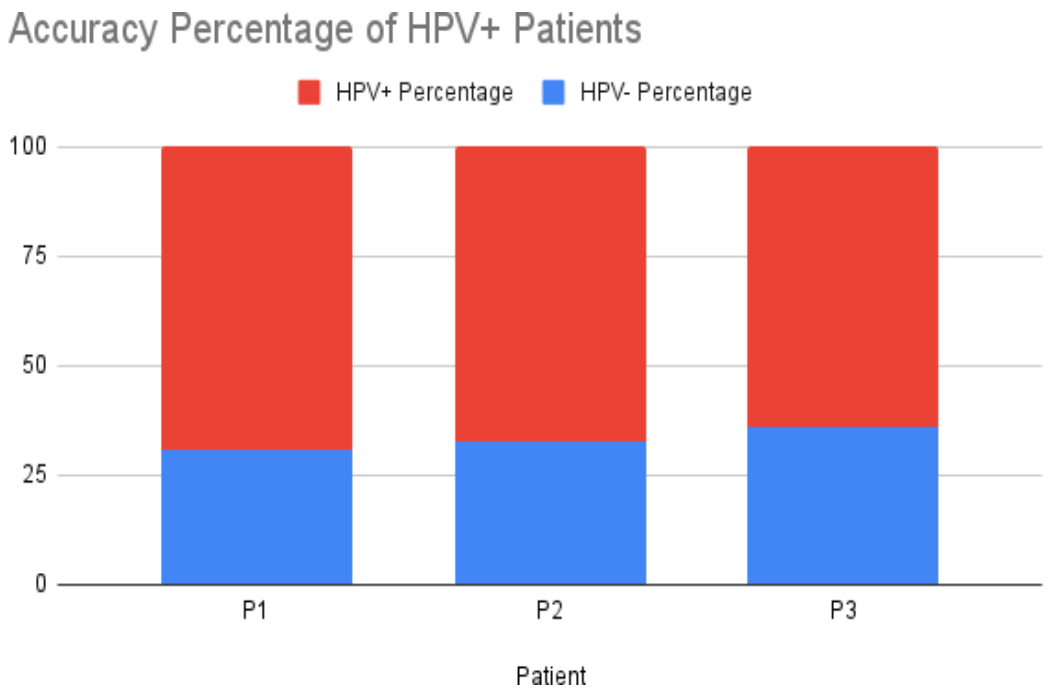

Figure 3.6: Diagrammatic representation of normal and diseased cells in HPV+ Patients

3.4 Data Visualization

Dimensionality reduction is the conversion of data from a high-dimensional to a low-dimensional space with the aim of maintaining the low-dimensional representation as close to the inherent dimension of the original data as feasible [32]. The sample integrity may result in incorrect classifications since sample tissues typically contain a mixture of both normal and disease-affected cells. In order to visualize the data points, we chose the 100 genes from the samples and plotted them using tsne and Umap in both 2D and 3D.

a) **t-SNE**:- Large datasets can be visualized using the dimensionality reduction technique known as t-Distributed Stochastic Neighbor Embedding (t-SNE). Each data point is assigned a place on a two- or three-dimensional map to represent the data [33]. Figure 3.8 3.8 below depicts the t-SNE 2-Dimensional and 3-Dimensional visualization. As we can see, the 2d graph shows a clear separation between both the classes of Diseased (HNSCC) cells and Normal (Healthy) cells. Whereas in the 3-dimensional representation, we can see some overlap between the cells. As some of thedata points seem to be merged with each other.

We can see that the classifications of diseased (HNSCC) and normal (healthy) cells are

clearly separated in the 2D graph in figure3.7. However, we can observe some cell overlapping in the three-dimensional depiction in figure 3.8. Since some of the data points appear to have been combined with one another.

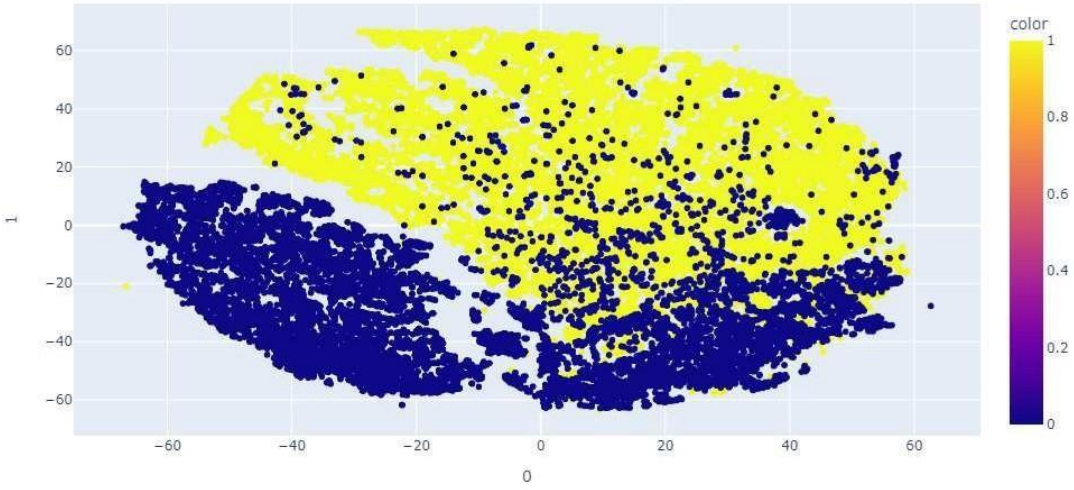

Figure 3.7: 2D visualization using tsne of both classes

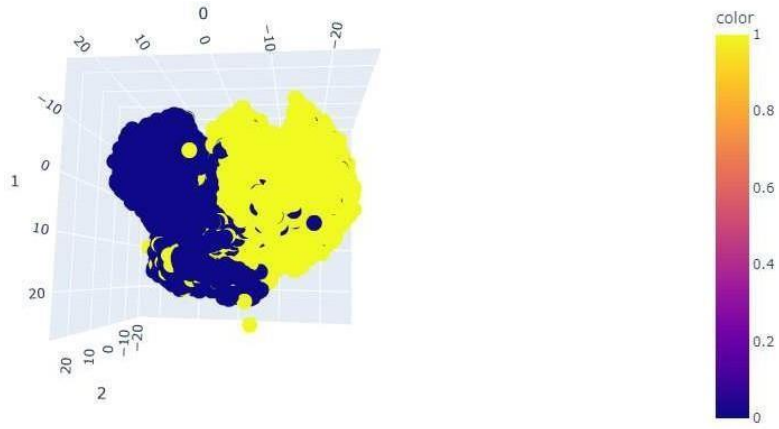

Figure 3.8: 3D visualization using tsne of both classes

b) **UMAP:-** The abbreviation UMAP stands for uniform manifold approximation and projection, which is a method of dimension reduction, like t-SNE, that could be applied to both general non-linear dimension reduction and visualizations. Despite having certain advantages over t-SNE in terms of separating batch effects, identifying pre-defined biological groups, and exposing in-depth clusters in two-dimensional space,

UMAP is superior to PCA and MDS. As it shows biological traits and clinical importance, therefore UMAP's sample clustering is important [34]. The dataset on the 100 selected genes has been represented in 2D and 3D using UMAP in figure 3.93.10. The data is clearly separated depending on the 100 genes that have been chosen as potential biomarkers.

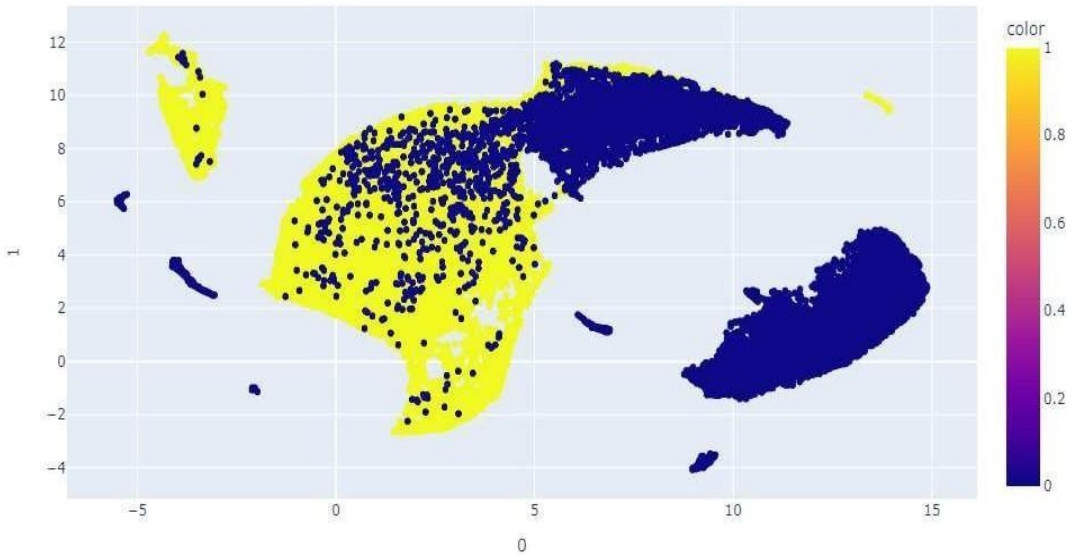

Figure 3.9: 2D visualization using umap of both classes

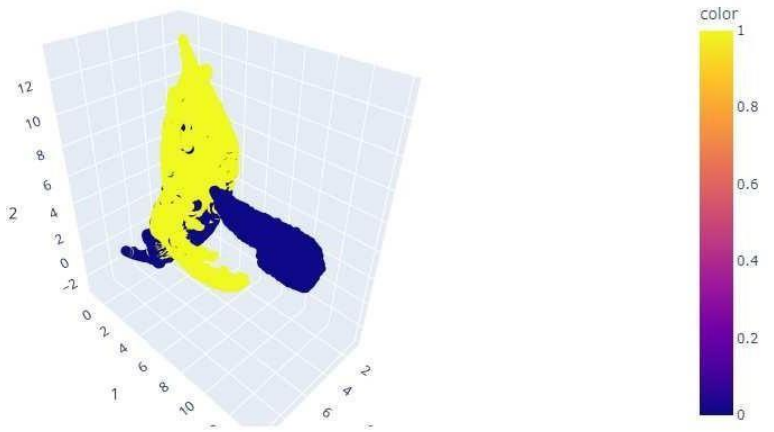

Figure 3.10: 3D visualization using umap of both classes

3.5 Gene Ontology

A systematic representation of a collection of information within a certain area is called an ontology. In most cases, ontologies are made up of a number of classes (or words or concepts) and the relations that exist between them. The Gene Ontology (GO) encapsulates our understanding of the biological world in three ways: Molecular Function, Cellular Component, and Biological Process [35, 36]. 100 genes that may serve as potential biomarkers of HNSCC were retrieved once mRMR analysis was complete. On these 100 retrieved genes, we next ran Gene Ontology (GO) Enrichment Analysis to map the biological functions of the chosen genes. The findings of the Go enrichment analysis are displayed in Table 3.5. We see that the majority of genes have a role in the catalytic and binding activities of many metabolic processes. ATP-dependent activity, molecular function regulator, molecular transducer, structural molecule activity, translation regulator activity, transcription regulator, and transporter activity are additional activities linked to the reported genes. A table3.5 of gene ontology is shown below:

GO Term	Activity	Genes
(GO:0140657)	ATP-dependent activity	ABCA8
(GO:0005488)	binding	"UBC",”ADAR’,’ICAM3’,’TXNIP’,’FXYD1’,’TIGIT’,’EEF1A1’,’RPLP1’ ’LIMD2’,’PSME1’,’RPS26’,’EZR’,’HLA-C’,’RHOG’,’FIGF’,’LTBP4’,’OAZ1’ ’B2M’,’RAC2’,’TIMP3’,’FHL1’,’UBB’,’HLA-B’,’BMP4’,’ARPC3’,’RELB’ ’EBF1’,’RHOH’,’ADH1B’,’RPS19’,’IL2RG’,’GSN’,’SOD3’,’GDF10’ ’SPARCL1’,’NOVA1’,’CREM’,’REL’,’HLA-A’
(GO:0003824)	catalytic activity	CYBRD1’, ’DUSP4’,’ADAR’,’PIM3’,’EEF1A1’,’RPLP1’,’PSME1’,’ISG20’,’ATP5E’,’RHOG’ ’OAZ1’,’RAC2’,’TIMP3’,’PTGIS’,’ABCA8’,’RHOH’,’ADH1B’,’SOD3’,’PLPP3’
(GO:0098772)	molecular function regulator	FXYD1’,’RPLP1’, ’PSME1’, ’FIGF’, ’OAZ1’, ’TIMP3’, ’BMP4’, ’GDF10’
(GO:0060089)	molecular transducer activity	FIGF’,’BMP4’, ’IL2RG’, ’GDF10’
(GO:0005198)	structural molecule activity	RPS15A ’, ’RPLP1’, ’RPS26’, ’RPS11’, ’RPS15’, ’RPS19’
(GO:0140110)	transcription regulator activity	RELB ’,’EBF1’, ’CREM’, ’REL’
(GO:0045182)	translation regulator activity	EEF1A1
(GO:0005215)	transporter activity	FXYD1’, ’ATP5E’, ’AQP7’, ’SLC16A3’, ’ABCA8’

Table 3.5: Table representing Gene ontology terms and their activity

CHAPTER 4

MODEL PACKAGING

For the scientific community and flexibility, we have developed a python package that can be used by anyone across the world using python environment. We also provided the web link for more details (https://webs.iiitd.edu.in/raghava/hnscpred/). Also the python package was uploaded on https://www.pypi.org and the package was named as "HNSCPred" for widespread usage. This package intends to categorize Normal Control (NC) patients and Head and Neck Cancer (HNSCC) patients from their single-cell RNAseq data using an artificial neural network (Deep Learning) model. The package uses 10x single-cell genomics data as input and a well-trained algorithm to determine whether the patient is diseased or healthy. The top 100 features that have the potential to serve as promising biomarkers in the classification and prediction of Normal and Diseased patients were discovered using the great feature selection approach known as mRMR (Minimum Redundancy Maximum Relevance). Additionally, individuals who were classified as diseased were further divided into HPV+ and HPV- categories.

4.1 Basic Package Structure:

Packages are a grouping of one or more modules. Typically, they are organized as a directory (the package) that contains one or more. modules in py file and/or subdirectory (also called sub packages). Python receives instructions from a specific file called init .py to interpret a directory as a package. The package's name has to be distinctive in order to avoid confusion with other users who have submitted files with the same name in the past. Ensure that the project's name matches the directory or folder containing the Python scripts. For new users, this makes setting and installation simpler.

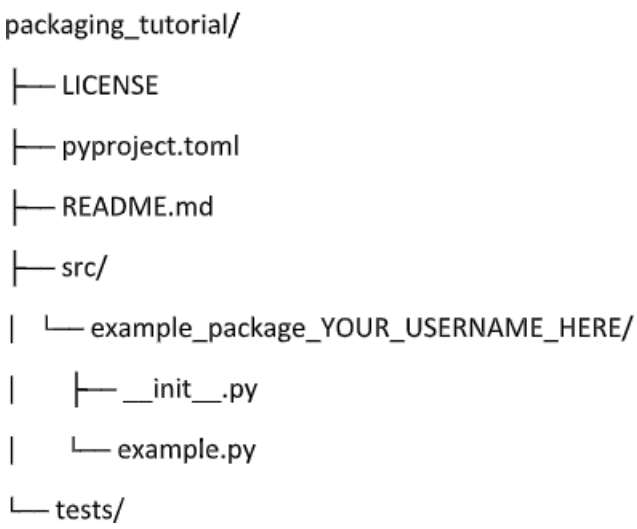

Figure 4.1: Package directory structure example

4.2 Package Requirements

For the package to be processed and installed quickly on any user machine, the package directory must also contain a number of crucial configuration files in addition to the Python script. The following additional files are necessary:

- Setup.py
- Manifest.in
- README.md
- LICENSE.txt
- Example files
- test files

All of the crucial configuration information and requirements for the programme installation are contained in the setup.py file. It includes information Version package metadata list of packages to include, list of other files to include, list of dependencies, list of extensions to be compiled etc. While, when we construct a Python package by default, just a limited set of files are included in the source distribution, the Manifest.in file is necessary to include extra critical files in our source distribution. Therefore, a MANIFEST.in file with all folder names and codes is necessary in order to incorporate more files and enable the system to identify them. For other users to use as a manual for

the project, the README.md file offers information on the project and a full description of it. Similarly, a LICENSE.txt file is a text file which is required for the copyrights, licenses, and restrictions.

4.3 Packaging process

- Converting to wheel file:
  After preparation of the package directory with all the requirements. The package needs to be converted into a .tar file and .whl (wheel) file with the below commands.

```
$ python3 setup.py sdist bdist_wheel
```

Figure 4.2: Wheel file creation code

- Uploading to PyPi:
  This command will create a wheel file for the source distribution of the pack- age and a tar file containing all the files of the package in the tar format. After preparation of the wheel file, you need to have twine installed which is required to upload the whl and tar file on www.pypi.org. The command file used is shown below. It will require the PyPI credential to log in and upload to the PyPI account.

```
$ twine upload dist/*
```

Figure 4.3: Upload on PyPI using twine

4.4 Package Details

4.4.1 HNSCPred

Using single-cell RNA seq data and computational methods, it is possible to predict which patients will develop HNSCC. From single-cell RNA seq data, this tool categorizes patients into Normal Control (NC) and Head and Neck Cancer (HNSCC) groups using an artificial neural network (Deep Learning) model. By using a highly trained algorithm and 10x single cell genomics data as input, it can determine if a patient is a healthy or diseased patient. The top 100 features for classification that serve as promising biomarkers in the classification and prediction of Normal and Diseased patients which were discovered using the outstanding feature selection approach knownas mRMR (Minimum Redundancy Maximum Relevance).

4.4.2 Installation

The user can download the package on any python environment with Python 3.0 or higher with the following command. It can also be found directly on www.pypi.org site. The screenshot of the PyPI is shown below.

```
pip install HNSCPred
```

Figure 4.4: pip install command to install package

Also in case the tool is previously installed, the tool can be upgraded to the latest version using the command.

```
pip3 install --upgrade HNSCPred
```

Figure 4.5: pip upgrade command to upgrade package

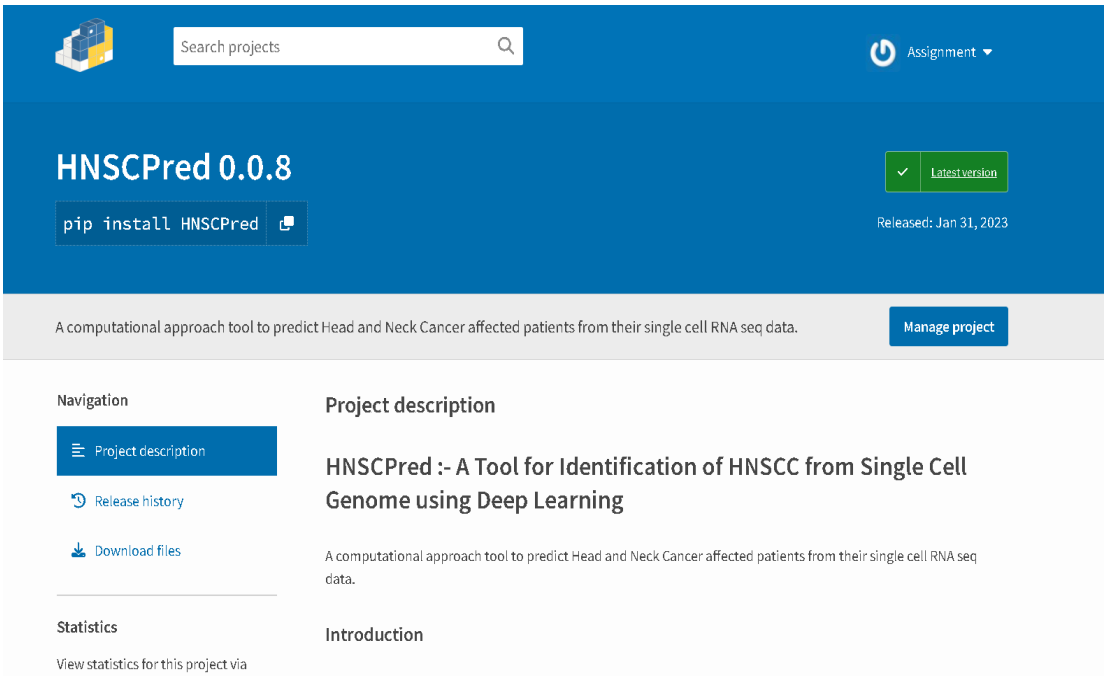

Figure 4.6: PyPI screenshot of package

4.4.3 Importing File

After installation of the HNSCPred package in your python environment, import the library using the below code. The HNSCPred comes with one inbuilt module. Please import the modules in your python environment before executing the code below.

```
import HNSCPred
```

Figure 4.7: Python code to import package

```
from HNSCPred import Validation
```

Figure 4.8: Python code to import Validation Module

4.4.4 Input

The input file should be a data frame in which the columns should be features (genes) and the rows should be cells. The file should contain the read count data of each gene in each cell. Please make sure that your single-cell data file is prepared in the above example.csv format. And the file should also contain the read count data for the selected 100 genes at https://webs.iiitd.edu.in/raghava/hnscpred/.

	Unnamed: 0	RP11-34P13.7	FO538757.2	AP006222.2	RP4-669L17.10	RP11-206L10.9	FAM87B	LINC00115	FAM41C	SAMD11	...	AC145212.2	AC011043
0	AAACGGGCATGACGGA.1	0	1	0	0	0	0	0	0	0	...	0	
1	AAAGATGAGCAGACTG.1	0	3	0	0	0	0	0	0	0	...	0	
2	AAAGATGAGTGTACTC.1	0	0	0	0	0	0	0	0	1	...	0	
3	AAAGATGCACTCTGTC.1	0	2	3	0	0	0	0	0	0	...	0	
4	AAAGCAAAGACAGGCT.1	0	0	0	0	0	0	0	0	0	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
603	TTTGGTTCACGAGGTA.1	0	0	0	0	0	0	0	0	0	...	0	
604	TTTGGTTGTTGTTTGG.1	0	1	1	0	0	0	0	0	0	...	0	
605	TTTGGTTTCGGAATCT.1	0	0	0	0	0	0	0	0	0	...	0	
606	TTTGTCAGTGCACCAC.1	0	2	0	0	0	0	0	0	0	...	0	
607	TTTGTCAGTGTTTGGT.1	0	1	2	0	0	0	0	0	0	...	0	

Figure 4.9: Example of input file

4.4.5 Demo

The demo to run the python package is shown in the figure below:

```
#Importing Inbuilt Module
from HNSCPred import Validation
```

```
#Reading patient single cell data (Diseased (HPV+ and HPV-) and Normal)
df1=pd.read_csv('/home/akankshaj/new CA data/Patient_wise_normal_dis/disease_p/Test_patient_disease/P4.csv')
df2=pd.read_csv('/home/akankshaj/new CA data/Patient_wise_normal_dis/disease_p/Test_patient_disease/P84.csv')
df3=pd.read_csv('/home/akankshaj/new CA data/Patient_wise_normal_dis/normal_p/Test_patient_normal/P38.csv')
```

```
#Predicting Results of Patient 1
Validation.predict(df1)
```

Figure 4.10: Code Demo

4.4.6 Output

The output of the file can be obtained by the code, as shown below. It will display the patient diagnosis, i.e., Diseased or Healthy, with the amount of Diseased or healthy cells found in the patient. Also classified diseased patients into sub-categories, i.e., HPV+ and HPV-.

```
#Predicting Results of Patient 1
Validation.predict(df1)
```

```
19/19 [==============================] - 0s 1ms/step
HNSCC patient detected, 99.3421052631579 percentage diseased cells detected
19/19 [==============================] - 0s 1ms/step
HNSCC HPV- patient detected, 100.0 percentage HPV- cells classified
```

```
#Predicting Results of Patient 2
Validation.predict(df2)
```

```
26/26 [==============================] - 0s 1ms/step
HNSCC patient detected, 98.3132530120482 percentage diseased cells detected
26/26 [==============================] - 0s 1ms/step
HNSCC HPV+ patient detected, 90.72289156626506 percentage HPV+ cells classified
```

```
#Predicting Results of Patient 3
Validation.predict(df3)
```

```
30/30 [==============================] - 0s 1ms/step
Normal patient detected, 44.164037854889585 percentage normal cells detected
```

Figure 4.11: Output

CHAPTER 5

DISCUSSION

One of the heterogeneous diseases is head and neck squamous cell carcinoma (HNSCC) which affects the head and neck region, namely the oral cavity, paranasal sinuses, larynx, nasal cavity, hypopharynx, and oropharynx. It is described by malignant and uncontrollable cell proliferation in these locations [37]. A biomarker is described as "a biological molecule present in the blood, other body fluids, as well as in tissues, that serves as a sign of a normal or aberrant process, a condition, or a disease" by the National Cancer Institute (NCI). To determine how effectively the body will react to an illness or condition medication, a biomarker could well be utilized [37]. Throughout the cancer diagnosis, biomarkers serve at least four crucial clinical functions, which include aiding in the diagnosis of cancer, predicting likely survival rates, assisting in the selection for a particular treatment depending on which patients are often the probable response, and determining at which dose range the drug may be most effective[38]. It can be challenging to make solid judgments from biomarker analysis data because of their well-known inconsistency [39]. The same biomarker may unmistakably reveal favorable relationships in other research while displaying negative or inconsistent results in other investigations—for instance, CCND1 [40], cMET [41], p16 [41, 42, 43], EGFR [44], and ERCC1 [45]. There are a number of possible explanations for these inconsistencies, such as limited sample sizes with insufficient controls, diverse research populations with factual clinical discrepancies as well as different therapeutic methods. Also, variability in the biomarker assay, such as various technological platforms used during identification and quantification and variations in the biomarker source. Additionally, there is a possibility of false positivity or false negativity with biomarker instability with other reasons such as various statistical testing techniques and methodological variations among research, such as comparing mRNA expression to protein expression [46, 39, 47, 37].

This study aims to find out a set of potential biomarkers from single-cell genomic data of head and neck cancer patients that can classify HNSCC and Normal patients with high accuracy. These biomarkers could aid in the early diagnosis and screening of

HNSCC. The biomarkers are shown in table 3.1. We have compiled multiple patients' single-cell data of HNSCC and NC patients, and the mRMR method was used to find out a subset of 100 features (genes) that could classify the samples with high accuracy.

The gene PLAC9's overexpression has been reported in connection with the inhibition of cell growth regulation and has also been reported in connection with cancers such as ovarian cancer and breast cancers as prognostic biomarkers [48]. Also, the ubiquitin UBB gene, when downregulated, inhibits the proliferation and radio resistance of cancer cells [49]. Gene 'ACKR1', along with other 3 genes in a study, was reported to be downregulated in HNSCC patients, which was co-related with poor prognosis ($p<0.05$) [50]. Also, gene 'AQP7', which is involved in physiologically functional cell migration, was upregulated in MSR of patients with Ten– tumors [51]. Whereas, geneFXYD1 was reported to be downregulated in the cancer samples, while FXYD4 and FXYD5 were overexpressed ($P<0.05$, fold change>1.5) [52]. In a study on cancer cells it was observed that BTG1 gene overexpression was linked to tumor growth or lung metastasis, inhibited proliferation, induced differentiation in different types of cancer cells [53]. Also mutations occurring in different genes, including B2M, CDKN2A, is found to show relation to the occurrence and development of tumors in Head and neck cancer patients [54].

Genes such as MFAP4, CD37, CXCL12, ADH1B, SOD3, SCARA5, ANGPTL1, FHL1, F10, CXCR4, MEG3, TXNIP, GDF10, and ABI3BP are shown to be downregulated because they operate as potential tumor suppressor genes, inhibiting tumor cell proliferation, invasion, and migration while also promoting apoptosis [55]. By control-ling the expression of miR-421 and E-cadherin, MEG3 long-encoding RNA inhibits the development of head and neck squamous cell carcinoma. However, additional research into MEG3's downstream mechanism in controlling the molecular process of epithelial-mesenchymal transformation (EMT) in head and neck squamous cell carcinoma (HN-SCC) development is required [56]. Growth differentiation factor-10 (GDF10), also known as BMP3b, is a tumor suppressor that belongs to the transforming growth factor-b (TGF-b) superfamily [57]. CIB1[58, 59], PIM3 [60], SLC16A3 [61], VOPP1 [62],
BMP4 [63], TIGIT [64], ADAR [65, 66], and LRRN4CL [67] are studied as upreguated genes. A complex that is important in the keratinocyte-intrinsic immune response to human papillomaviruses (-HPVs) is formed when CIB1 interacts with the EVER1, and

EVER2 proteins [58, 59]. It has been observed that nearly all primary HNSCCs express at least one PIM kinase member at high levels [60]. Immunological checkpoint T cell immunoreceptor with immunoglobulin and ITIM domain (TIGIT) is essential for immune suppression. However, it has a connection to genetics and epigenetics, and a role in tumor immunity [64]. The transforming growth factor (TGF) superfamily includes extracellular signaling molecules known as bone morphogenetic proteins (BMPs), which are known to control cell proliferation, differentiation, and motility, particularly during development [63]. Functional research shows that, particularly in HNSCC cancer, has connected BMP4 to the encouragement of cell migration and the suppression of cell proliferation [63].

Overall most of the genes which were obtained from our study have been reported as promising candidate for biomarkers in various studies [50, 51, 52, 53, 54, 55, 60]. Still, some genes have not yet been reported in connection with Head and Neck cancer. These genes may require further investigation and study and may act as novel findings which could help diagnose patients with Head and neck cancer. The python package is also created using a computational approach to predict Head and Neck Cancer affected patients from their single-cell RNA seq data (https://pypi.org/project/HNSCPred/). With the use of an artificial neural network (Deep Learning) model, this package aims to classify Normal Control (NC) patients and Head and Neck Cancer (HNSCC) patients from their single-cell RNA seq data. The tool employs a trained algorithm and 10x single-cell genomics data as input to decide if the patient is diseased or healthy. Using the excellent feature selection method known as mRMR, the top 100 features that have the ability to act as promising biomarkers in the classification and prediction of Normal and Diseased patients were found (Minimum Redundancy Maximum Relevance). Additionally, groups of people who were labelled as diseased were separated into HPV+ and HPV- categories. We also provided the web link for more details (https://webs.iiitd.edu.in/raghava/hnscpred/).

CHAPTER 6

CONCLUSION

To categorize Normal Control (NC) cells and HNSCC disease cells from their single-cell RNA seq data, we employed a variety of machine learning models, including an ANN deep learning model. We also further categorized diseased patients into HPV+ and HPV-. In this study, patient-wise analysis is also applied to categorize the samples. We trained the model with two datasets (GSE181919 and GSE139324) and evaluated that the Deep Learning model performs better against other Machine Learning models. We subsequently compared the model's performance between two datasets and found model trained on dataset GSE181919 performed better. The datasets were originally quite extensive and had a significant number of features. During the pre- processing step, the feature count was decreased to a shallow level (approx 3500). One of the feature selection techniques known as mRMR was used to obtain a limited set of features which could be helpful in categorizing the samples because many characteristics were co-related and duplicated. The top 100 features with the least amount of redundancy and the most relevance were extracted from these 3500 features using mRMR. Further, 100 genes (features) separated the HNSCC patients from NC patients with an accuracy of 92 percent, an AUC-ROC of 0.91 in internal validation, and an accuracy of 74 percent, an AUC-ROC of 0.66 in external validation. Whereas in the case of HPV classification, the metrics obtained were, AUC-ROC 0.83 and 84 percent accuracy in internal validation and 0.48 AUC-ROC and 69 percent accuracy in external validation. For the detection and categorization of biomarkers, ANN has proven to be an effective technique among all machine learning models. In order to help the scientific community, we ended up creating a Python package called "HNSCPred" based on the aforementioned work (https://webs.iiitd.edu.in/raghava/hnscpred/). To fully understand how the discovered genes impact and contribute to the progression of HNSCC Disease, further clinical investigations on these genes are necessary.

REFERENCES

[1] M. D. Mody, J. W. Rocco, S. S. Yom, R. I. Haddad, and N. F. Saba, "Head and neck cancer," *The Lancet*, 2021.

[2] H. Mahmood, M. Shaban, N. Rajpoot, and S. A. Khurram, "Artificial intelligence-based methods in head and neck cancer diagnosis: An overview," *British journal of cancer*, vol. 124, no. 12, pp. 1934–1940, 2021.

[3] D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis, "Head and neck squamous cell carcinoma," *Nature reviews Disease primers*, vol. 6, no. 1, pp. 1–22, 2020.

[4] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[5] N. Basheeth and N. Patil, "Biomarkers in head and neck cancer an update," *Indian Journal of Otolaryngology and Head & Neck Surgery*, vol. 71, no. 1, pp. 1002–1011, 2019.

[6] R. Radpour and F. Forouharkhou, "Single-cell analysis of tumors: Creating new value for molecular biomarker discovery of cancer stem cells and tumor-infiltrating immune cells," *World journal of stem cells*, vol. 10, no. 11, p. 160, 2018.

[7] J. Eberwine, J.-Y. Sul, T. Bartfai, and J. Kim, "The promise of single-cell sequencing," *Nature methods*, vol. 11, no. 1, pp. 25–27, 2014.

[8] A.-E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, "Single-cell rna-seq: advances and future challenges," *Nucleic acids research*, vol. 42, no. 14, pp. 8845–8860, 2014.

[9] C. Gawad, W. Koh, and S. R. Quake, "Single-cell genome sequencing: current state of the science," *Nature Reviews Genetics*, vol. 17, no. 3, pp. 175–188, 2016.

[10] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, 2015.

[11] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, *et al.*, "Highly paral-lel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.

[12] S. Liu and C. Trapnell, "Single-cell transcriptome sequencing: recent advances and remaining challenges," *F1000Research*, vol. 5, 2016.

[13] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Reviews Genetics*, vol. 16, no. 3, pp. 133–145, 2015.

[14] A. R. Cillo, C. H. Kürten, T. Tabib, Z. Qi, S. Onkar, T. Wang, A. Liu, U. Duvvuri, S. Kim, R. J. Soose, *et al.*, "Immune landscape of viral-and carcinogen-driven head and neck cancer," *Immunity*, vol. 52, no. 1, pp. 183–199, 2020.

[15] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, pp. 523–528, 2003.

[16] B. Niu, G. Huang, L. Zheng, X. Wang, F. Chen, Y. Zhang, and T. Huang, "Prediction of substrate-enzyme-product interaction based on molecular descriptors and physicochemical properties," *BioMed research international*, vol. 2013, 2013.

[17] S.-C. Wang, "Artificial neural network," in *Interdisciplinary computing in java programming*, pp. 81–100, Springer, 2003.

[18] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[19] R. E. Wright, "Logistic regression.," 1995.

[20] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[21] J. R. Quinlan, "Learning decision tree classifiers," *ACM Computing Surveys (CSUR)*, vol. 28, no. 1, pp. 71–72, 1996.

[22] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[23] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[24] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[26] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, "K-nearest neighbor classification," in *Data mining in agriculture*, pp. 83–106, Springer, 2009.

[27] L. Peng, X. W. Bian, D. K. Li, C. Xu, G. M. Wang, Q. Y. Xia, and Q. Xiong, "Large-scale rna-seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 tcga cancer types," *Scientific reports*, vol. 5, no. 1, pp. 1–18, 2015.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[29] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, no. 1, pp. 1–5, 2018.

[30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[31] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[32] L. Van Der Maaten, E. Postma, J. Van den Herik, *et al.*, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[34] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell, "Dimensionality reduction by umap to visualize physical and genetic interactions," *Nature communications*, vol. 11, no. 1, pp. 1–6, 2020.

[35] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[36] "The gene ontology resource: enriching a gold mine," *Nucleic acids research*, vol. 49, no. D1, pp. D325–D334, 2021.

[37] J. C.-H. Hsieh, H.-M. Wang, M.-H. Wu, K.-P. Chang, P.-H. Chang, C.-T. Liao, and C.-T. Liau, "Review of emerging biomarkers in head and neck squamous cell carcinoma in the era of immunotherapy and targeted therapy," *Head & Neck*, vol. 41, pp. 19–45, 2019.

[38] C. L. Sawyers, "The cancer biomarker problem," *Nature*, vol. 452, no. 7187, pp. 548–552, 2008.

[39] J. P. Rodrigo, L. A. García, S. Ramos, P. S. Lazo, and C. Suárez, "Ems1 gene amplification correlates with poor prognosis in squamous cell carcinomas of the head and neck," *Clinical cancer research*, vol. 6, no. 8, pp. 3177–3182, 2000.

[40] A. A. da Costa, F. D. Costa, D. V. Araújo, M. P. G. Camandaroba, V. H. F. de Jesus, A. Oliveira, A. C. F. Alves, C. Stecca, L. Machado, A. C. F. de Oliveira, *et al.*, "The roles of pten, cmet, and p16 in resistance to cetuximab in head and neck squamous cell carcinoma," *Medical oncology*, vol. 36, no. 1, pp. 1–9, 2019.

[41] L. Satgunaseelan, N. Chia, H. Suh, S. Virk, B. Ashford, T. Lum, M. Ranson, J. Clark, and R. Gupta, "p16 expression in cutaneous squamous cell carcinoma of the head and neck is not associated with integration of high risk hpv dna or prognosis," *Pathology*, vol. 49, no. 5, pp. 494–498, 2017.

[42] B. Pajares, L. Perez-Villa, J. M. Trigo, M. Toledo, M. Alvarez, B. Jimenez, J. Medina, V. de Luque, J. Jerez, and E. Alba, "Concurrent radiotherapy plus epidermal growth factor receptor inhibitors in patients with human papillomavirus-related head and neck cancer," *Clinical and Translational Oncology*, vol. 16, no. 4, pp. 418–424, 2014.

[43] M. Lundberg, I. Leivo, K. Saarilahti, A. A. Mäkitie, and P. S. Mattila, "Transforming growth factor beta 1 genotype and p16 as prognostic factors in head and neck squamous cell carcinoma," *Acta oto-laryngologica*, vol. 132, no. 9, pp. 1006–1012, 2012.

[44] V. Bišof, M. Zajc Petranović, Z. Rakušić, K. R. Samardžić, and A. Juretić, "The prognostic and predictive value of excision repair cross-complementation group 1 (ercc1) protein in 1288 patients with head and neck squamous cell carcinoma treated with platinum-based therapy: a meta-analysis," *European Archives of Oto-Rhino-Laryngology*, vol. 273, no. 9, pp. 2305–2317, 2016.

[45] G. Poste, "Bring on the biomarkers," *Nature*, vol. 469, no. 7329, pp. 156–157, 2011.

[46] R. Simon, "Clinical trials for predictive medicine: new challenges and paradigms," *Clinical trials*, vol. 7, no. 5, pp. 516–524, 2010.

[47] K. Y. Kim, L. M. McShane, and B. A. Conley, "Designing biomarker studies for head and neck cancer," *Head & neck*, vol. 36, no. 7, pp. 1069–1075, 2014.

[48] C. Ouyang, Y.-Z. Pu, X.-H. Qin, J. Shen, Q.-H. Liu, L. Ma, and L. Xue, "Placenta-specific 9, a putative secretory protein, induces g2/m arrest and inhibits the proliferation of human embryonic hepatic cells," *Bioscience reports*, vol. 38, no. 6, 2018.

[49] Y. Tang, Y. Geng, J. Luo, W. Shen, W. Zhu, C. Meng, M. Li, X. Zhou, S. Zhang, and J. Cao, "Downregulation of ubiquitin inhibits the proliferation and radioresistance of non-small cell lung cancer cells in vitro and in vivo," *Scientific reports*, vol. 5, no. 1, pp. 1–12, 2015.

[50] H. Liu, G. Hei, L. Zhang, Y. Jiang, and H. Lu, "Identification of a novel cerna network related to prognosis and immunity in hnscc based on integrated bioinformatic investigation," *Scientific Reports*, vol. 12, no. 1, pp. 1–15, 2022.

[51] V. Zivicova, P. Gal, A. Mifkova, S. Novak, H. Kaltner, M. Kolar, H. Strnad, J. Sachova, M. Hradilova, M. Chovanec, *et al.*, "Detection of distinct changes in gene-expression profiles in specimens of tumors and transition zones of tenascin-positive/-negative head and neck squamous cell carcinoma," *Anticancer Research*, vol. 38, no. 3, pp. 1279–1290, 2018.

[52] M. Jin, H. Zhang, J. Yang, Z. Zheng, and K. Liu, "Expression mode and prognostic value of fxyd family members in colon cancer," *Aging (Albany NY)*, vol. 13, no. 14, p. 18404, 2021.

[53] S. Zhao, H. Xue, C.-l. Hao, H.-m. Jiang, and H.-c. Zheng, "Btg1 overexpression might promote invasion and metastasis of colorectal cancer via decreasing adhesion and inducing epithelial–mesenchymal transition," *Frontiers in Oncology*, vol. 10, p. 598192, 2020.

[54] J. Wang, X. Chen, Y. Tian, G. Zhu, Y. Qin, X. Chen, L. Pi, M. Wei, G. Liu, Z. Li, *et al.*, "Six-gene signature for predicting survival in patients with head and neck squamous cell carcinoma," *Aging (Albany NY)*, vol. 12, no. 1, p. 767, 2020.

[55] Y. Sun, Q. Zhang, L. Yao, S. Wang, and Z. Zhang, "Comprehensive analysis reveals novel gene signature in head and neck squamous cell carcinoma: predicting is associated with poor prognosis in patients," *Translational Cancer Research*, vol. 9, no. 10, p. 5882, 2020.

[56] Y. Ji, G. Feng, Y. Hou, Y. Yu, R. Wang, and H. Yuan, "Long noncoding rna meg3 decreases the growth of head and neck squamous cell carcinoma by regulating the expression of mir-421 and e-cadherin," *Cancer medicine*, vol. 9, no. 11, pp. 3954–3963, 2020.

[57] C.-W. Cheng, J.-R. Hsiao, C.-C. Fan, Y.-K. Lo, C.-Y. Tzen, L.-W. Wu, W.-Y. Fang, A.-J. Cheng, C.-H. Chen, I.-S. Chang, *et al.*, "Loss of gdf10/bmp3b as a prognostic marker collaborates with tgfbr3 to enhance chemotherapy resistance and epithelial-mesenchymal transition in oral squamous cell carcinoma," *Molecular carcinogenesis*, vol. 55, no. 5, pp. 499–513, 2016.

[58] L. D. Notarangelo, "Hpv: Cib1 is for ever and ever," *Journal of Experimental Medicine*, vol. 215, no. 9, pp. 2229–2231, 2018.

[59] S. J. De Jong, A. Créquer, I. Matos, D. Hum, V. Gunasekharan, L. Lorenzo, F. Jabot-Hanin, E. Imahorn, A. A. Arias, H. Vahidnezhad, *et al.*, "The hu- man cib1–ever1–ever2 complex governs keratinocyte-intrinsic immunity to β-papillomaviruses," *Journal of Experimental Medicine*, vol. 215, no. 9, pp. 2289–2310, 2018.

[60] T. R. Broutian, B. Jiang, J. Li, K. Akagi, S. Gui, Z. Zhou, W. Xiao, D. E. Symer, and M. L. Gillison, "Human papillomavirus insertions identify the pim family of serine/threonine kinases as targetable driver genes in head and neck squamous cell carcinoma," *Cancer letters*, vol. 476, pp. 23–33, 2020.

[61] S. Yu, Y. Wu, C. Li, Z. Qu, G. Lou, X. Guo, J. Ji, N. Li, M. Guo, M. Zhang, *et al.*, "Comprehensive analysis of the slc16a gene family in pancreatic cancer via integrated bioinformatics," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[62] A. S. Baras, A. Solomon, R. Davidson, and C. A. Moskaluk, "Loss of vopp1 overexpression in squamous carcinoma cells induces apoptosis through oxidative cellular injury," *Laboratory investigation*, vol. 91, no. 8, pp. 1170–1180, 2011.

[63] E.-L. Alarmo, H. Huhtala, T. Korhonen, L. Pylkkänen, K. Holli, T. Kuukasjärvi, S. Parkkila, and A. Kallioniemi, "Bone morphogenetic protein 4 expression in multiple normal and tumor tissues reveals its importance beyond development," *Modern Pathology*, vol. 26, no. 1, pp. 10–21, 2013.

[64] J. Wen, X. Mao, Q. Cheng, Z. Liu, and F. Liu, "A pan-cancer analysis revealing the role of tigit in tumor microenvironment," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[65] X.-x. Huo, S.-j. Wang, H. Song, H. Yu, M. Wang, H.-x. Gong, X.-t. Qiu, Y.-f. Zhu, J.-y. Zhang, *et al.*, "Roles of major rna adenosine modifications in head and neck squamous cell carcinoma," *Frontiers in Pharmacology*, vol. 12, 2021.

[66] D. Zheng and B. Tian, “Rna-binding proteins in regulation of alternative cleavage and polyadenylation,” *Systems Biology of RNA Binding Proteins*, pp. 97–127, 2014.

[67] F. Yang, L.-q. Zhou, H.-w. Yang, and Y.-j. Wang, “Nine-gene signature and nomogram for predicting survival in patients with head and neck squamous cell carcinoma,” *Frontiers in Genetics*, vol. 13, 2022.