HNSCC is squamous cell carcinomas arising from lesions in the head and neck region, including the oral cavity, tongue, nasal and paranasal cavity, pharynx, and larynx(Promoter).  HNSCC is among the top ten leading causes of cancer worldwide, with approximately 500,000 new cases diagnosed annually (Prognostic implication). The high prevalence and risk associated with HNSCC emphasize the importance of developing effective methods to assess cancer progression and predict patient prognosis.

Numerous studies have developed various biomarkers with prognostic value for HNSCC. Among all these biomarkers, gene signatures, particularly the Hypoxia Signature (HS) – a set of differentially expressed genes (DEGs) under hypoxic conditions within the tumour microenvironment – have emerged as a key method. The expression level of HS can be used to infer intratumoural hypoxia levels (Hypoxic Characteristic G), which is crucial for predicting clinical outcomes (Comparative analysis of tra). This signature-identification-cantered method also applied on protein expression data. It has been shown that proteomic biomarkers can function as predictor for cancer diagnosis and prognosis [(Tumor tissue hnRNP M)](https://www.sciencedirect.com/science/article/pii/S0009898120302862#b0020), and some proteomic signature has been developed on the Reverse phase protein array (RPPA) data (Prognostic implication).

Despite these advances, challenges remain, particularly in harnessing the full potential of multi-omics data due to its high dimensionality, which complicates the extraction of meaningful information from noises. AEs have been introduced as a solution to the dimensionality problem. However, existing work either rely on supervised method [A multi-omics supervised a], which may not be feasible when labels are unavailable, or are lacking application to protein expression data [Survival stratification for c], or do not focus specifically on HNSCC [AFExNet: An Adversari].

These research gap necessitates this research, aiming to determine whether protein expression data can provide insights into patient OS and develop an AE-based risk stratification model for accurate OS classification in HNSCC patients.

The structure of this paper is as follows: Section 2 reviews the related work in the field, providing context and highlighting the novelty of our approach. Section 3 delves into the experimental design, detailing the AERO-HNSCC architecture, data preparation procedures, validation metrics, and benchmarking strategies. Section 4 analyses the results, assessing the efficacy and potential implications of our model. Finally, Section 5 offers a discussion, reflecting on the findings and suggesting directions for future research.

The Related Work section of this paper delves into two pivotal methodologies that have significantly impacted the prognosis and classification HNSCC: SIC Methods and AE-Based Methods. Understanding these approaches illuminates the backdrop against which this research is situated.

1. *SIC Methods*

SIC methods have been cornerstone approaches in deciphering the complex molecular landscapes of cancers, including HNSCC. These methods focus on identifying specific gene signatures that correlate with disease outcomes, HSs especially. All SIC related paper reviewed is shown in table I.

As a conclusion of papers listed above, the development of HS can be generalized as a three-stage process for various tumour types:

* 1. **Identification of Hypoxia DEGs**  
     This initial step involves obtaining all hypoxia-related DEGs, which can be sourced through literature surveys, directly from databases, or identified using algorithms like K-mean or UHC.
  2. **Dimensionality Reduction on DEGs**

Subsequently, dimensionality reduction techniques such as LASSO, LASSO Cox Regression, or Random Forest are applied to these DEGs to pinpoint prognostic HSs.

* 1. **Prognostic Model Development**  
     The final step involves developing a prognostic model, which may take the form of a score or a more complex model incorporating additional features.

Specially, while all other works fucus on gene expression dataset, Wu et al.’s work following the identical three-stage process, but developed a proteomic signature on the TCGN-HNCC RPPA dataset, achieved a 0.779 of area under the curve of the corresponding receiver operating characteristic in the task of classify patients into high and low risk groups.

Due to the variance in tumour types, datasets used for HSs development, the result measurement and processes of HS identification, conducting comparative research on the performances of all different methods remains challenging. Noted the overlapping focus on HCC and LUAD, this research will fucus on the performance of 6 works developed HSs for these two tumours, respectively.

give me a example of Introduction to AE-Based Methods: Introduce Autoencoders (AEs), explaining their basic principle and why they represent a shift from traditional signature-identification methods:

Supervised:

Tan et al.’s work focus on the multi-omics data of pan-cancer, they construct isolated AEs for every omics data including DNA methylation, miRNA-Seq, RNA-seq, and protein expression. These AEs encoded data to a unfifom dimensionality and trained with labels, including overall survival (OS), disease-specific survival (DSS), progression-free interval (PFI) and disease-free interval (DFI). Their work achieved a AUC of 0.7830 for binary classification.

Mondol et al.’s work present a unsupervised pre-training and supervised fine-tuning adversarial auto-encoder and a method named ‘TopGene’ to find highly weighted genes from the latent space. Achieved precision of 0.8596 on sub-types classify. TPOT Optimised. Developed on RNA-Seq dataset of breast cancer, use sub-types as labels when fine-tuning.

Madhumita and Sushmita conduct a supervised feature selection post a spare AE, developed for GBM and clustering sub-types, therefore no AUC or precision reported.

Unsupervised:

Song et al.’s work focus on survival stratification for colorectal cancer, DNA methylation, RNA-seq, and miRNA-seq data are used. They achieved a C-index of 0.781 with a deep spare AE.

Ellen et al.’s work construct a single layer denoising zeros AE with input data the combination of mRNA, miRNA, DNA methylation, long non-coding RNA (lncRNA). The result is concordance indices (C‑indices) of 0.69 ± 0.03 for LUAD.

**Similarly,** Arafa et al. contributed a Reduced Noise Autoencoder, which is a 3 layer AE with Reduced Noise-Synthesis Minority Over Sampling Technique (***RN- SMOTE***) **achieved a 0.75 Precision on** colon cancer dataset, only genomic data.

no preunderstanding of gene or protein needed. Can take the advantage of dimensionality reduction and extend data to multi-omics data instead of single gene or protein expression. The process is universal for all types of cancers, as long as data prepared and the OS signal of the type of cancer is contained in the multi-omics data. More importantly, AE encoded data can capture the potential more complicated non-linear associations with the OS, beyond the SIG which can only capture the linear relationships between a certain biomarker and the OS.

B. AE-Based Methods

In contrast to Signature-Identification-Centred (SIC) Methods, AE-Based Methods offer a fresh perspective on data analysis in oncological research, particularly through the lens of multi-omics integration. This shift is due to the complexities of cancer that extend beyond linear biomarker associations, emphasizing the potential of non-linear data relationships and the holistic nature of biological systems. Current works can be divided into two paths: supervised and unsupervised.

Supervised Approaches:

Tan et al.'s work focused on pan-cancer multi-omics datasets, constructing individual autoencoders for each data type—ranging from DNA methylation to protein expression. By encoding these varied data types to a uniform dimensionality and training with distinct labels such as overall survival (OS) and disease-specific survival (DSS), their approach achieved a noteworthy AUC of 0.7830 for binary classification.

Similarly, Mondol et al. presented a blend of unsupervised pre-training with supervised fine-tuning through an adversarial autoencoder and a method named 'TopGene.' This approach identifies significant genes within the latent space, demonstrating high precision of 0.8596 in classifying sub-types of breast cancer.

Madhumita and Sushmita's work, although not quantified through AUC or precision metrics, focuses on the subtype clustering of glioblastoma multiforme (GBM). They contribute a method that performs supervised feature selection before training a sparse autoencoder.

Unsupervised Approaches:

On the unsupervised frontier, AE-Based Methods demonstrate their strength in survival stratification without the prerequisite of predefined labels. Song et al.’s research in colorectal cancer utilized DNA methylation, RNA-seq, and miRNA-seq data with a deep sparse AE, achieving a concordance index (C-index) of 0.781.

Ellen et al. constructed a single-layer denoising autoencoder that integrates mRNA, miRNA, DNA methylation, and long non-coding RNA data, yielding concordance indices (C indices) of 0.69 ± 0.03 for lung adenocarcinoma (LUAD).

Arafa et al.’s work with a Reduced Noise Autoencoder showcases the utility of AE in enhancing data quality through noise reduction. Utilizing a three-layer AE with Reduced Noise-Synthesis Minority Over Sampling Technique (RN-SMOTE), they achieved a precision of 0.75 on a colon cancer dataset using only genomic data.

Transition and Comparative Outlook:

While AE-Based Methods are forging new paths in cancer research, they stand in stark contrast to SIC Methods by offering a more flexible and comprehensive approach to data analysis. Characterized by their ability to handle high-dimensional data and applicable to various cancer types without the need for prior biological knowledge, these methods uncover complex, non-linear relationships that traditional methods might overlook. As this research transitions from the specificities of SIC Methods to the broad potential of AE-Based Methods, it becomes evident that the latter may offer a more integrative and universal framework for understanding the multifaceted nature of cancer progression and patient prognosis.

1. Supervised method contradict to the wish of being more generative by its request of labels, although some of their findings and insight is still helpful on the design of the architecture of AEs.
2. Among all *unsupervised* works, there is no research focused on HNSCC, or integrated protein expression, or the OS of HNSCC. this leaves us a gap to fill.

Start from a regular AE: input layer of size 468 (protein expression number) connect to 2 hidden layers of size 64 and 32 with relu as activation function. For the output layer a sigmoid is appled. Adam optimizer and MSE reconstruction loss. This AE is trained with 3 varations: bottleneck of 2,6,12.

Realising to capture the OS signal require higher dimensions, a wider deep AE(DAE) is designed. This DAE connect the input layer to three hidden layers of size 256,128,64 with other setting remains the same. This DAE is trained on bottleneck 12, 48, 36, 34, 32, 28, 24, and 18 and reach a record of precision when bottleneck is set to 18.

After then, we modify the initial AE to a spare AE(SAE) by applying l1 reularization on its hidden layers with a L1 penaty rate of 1e-6, seeing a noticeable improvement, we therefore also tried it on the DAE make it then a spare DAE(SDAE).

AERO- HNSCC is developed on the TCGN-HNSCC RPPA dataset, the overall process of data is described by fig 3. The data downloaded from TCGA is separated in 353 individual tsv files. We merged them into a single pandas dataframe for following processing. After carefully checking, we identified 18 columns of nan value and 1 column that miss in 280 cases, and 10 cases that missing 218 columns. These data are dropped due to unrecoverable percentage missing. In the remaining dataframe, left 141 cases that have 12 columns of missing value. Considering the amount of cases and the relatively controllable number of missing columns (12 to 468 in total), we decide to do imputation by integrate clinical data. We merged the clinical data with the RPPA dataset by the case\_submitter\_id, which is the unique id of each patient so that we can link the clinical information with the RPPA data. Then we group cases which have no missing values by the ajcc\_pathologic\_stage and calculate the median of all 12 columns for each group. Having the median value, we move back to the cases that these 12 columns are missing, and assign back the median value by groups. This target encoding like imputation enable us to fully explore the advantage of multi-omics data and conducting meaningful and effectual imputation by introducing new information.

Having the clinical data integrated, a second filtering is conduct for filter out 2 extra cases that have ambiguous OS, result our final data: 324 case with 468 protein expression columns, as shown in **Supplementary Fig. 1**

**Then it comes to the normalization of input features, i.e., the protein expression. As shown in fig 4, observed that the mean of protein expression varises from 2 to 0.001, it is curtail to conduct normalization for the reason that […]**

Employed Adam optimizer, attached the test set as validation data and tested after each epoch, set the checkpoint to save the epoch performed best on the test set. The model is trained with 100 epochs with a batch size of 8. The loss get for each AE is shown in **Supplementary Fig. 2.**

|  |  |
| --- | --- |
|  |  |
| Initial Architecture, 468 Protein Expressions Encode to Dimension of 6 | Initial Architecture, 456 Protein Expressions Encode to Dimension of 6 |
|  |  |
| Initial Architecture, 468 Protein Expressions Encode to Dimension of 2 | Initial Architecture, 456 Protein Expressions Encode to Dimension of 2 |
|  |  |
| Initial Architecture, 468 Protein Expressions Encode to Dimension of 12 | DAE, 468 Protein Expressions Encode to Dimension of 12 |
|  |  |
| DAE, 468 Protein Expressions Encode to Dimension of 48 | DAE, 468 Protein Expressions Encode to Dimension of 36 |
|  |  |
| DAE, 468 Protein Expressions Encode to Dimension of 3 | DAE, 468 Protein Expressions Encode to Dimension of 32 |
|  |  |
| DAE, 468 Protein Expressions Encode to Dimension of 28 | DAE, 468 Protein Expressions Encode to Dimension of 24 |
|  |  |
| **DAE, 468 Protein Expressions Encode to Dimension of 18** | SAE, 468 Protein Expressions Encode to Dimension of 12 |
|  |  |
| DSAE, 468 Protein Expressions Encode to Dimension of 18 |

The benchmark model includes three most relevant and state-of-the-art works: Brooks et al.’s work: they validated the SIG method developed gene signature for HNSCC, obtained a result of P = 0.5/0.2 in log rank test of two validation cohort dataset and claim HS not independently prognostic.

Another work is contributed by Donglei et al., they validated a SIG method developed [Proteomic](https://www.sciencedirect.com/topics/medicine-and-dentistry/proteomics) signature for HNSCC, achived a AUC of 0.779. they “As expected, prognostic risk model based on risk proteins function as a predictor for prognosis in HNSCC (HR:1.296 and P < 0.05). The area under the curve of the corresponding receiver operating characteristic (ROC) curve for prognostic risk model based on risk proteins is 0.779. ”

The last work selected is [Kaiwen](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1114-3#auth-Kaiwen-Tan-Aff1) et al.’s. Though not specially focus on HNSCC, their work contribute a pan-cancer AE-based method for applied on OS perdiction. Their *supervised approach achieved a AUC of 0.7830.*

As the primary step of the objective of this research, as mentioned in the introduction, we wish to firstly exam if protein expression data contains HNSCC OS insights. Therefore, after obtained the cleaned RPPA data, a exploration classification is conducted by SVM with linear kernel. The result accuracy for this binary classification is 0.58 which is significantly above 50%. This good signal lead us to the following research.

Then we trained first 4 AEs, initial architecture encoded all 468 protein to dimension of 6 and 2, with and without the imputation columns. By include the imputation, an 8.51% and 4.08% increase on accuracy is observed, therefore verify the validation of imputation. However, the best result obtained by the imputation data encoded to 6 dimension is 0.59, not obvious improvement comparing with the raw RPPA data. Doing visualization by apply PCA with component of 3 and 2 on the 6 dimension result, and plot the encoded 2 dimension data, as shown in fig 5, there is no clear boundaries between two risk groups. Observe the improvement when go up to 6 dimension, we realize the AE may need more dimension to caputure the pattern. Therefore another 12 dimension AE is trained. Here we are getting close to the upper hidden layer size, to discovery the performance could be further improved by wider and deeper AE, we design the DAE and trained 8 variations with different bottolenecks, going down from 48, we reach a best record when bottoleneck is 18, which have a percision of 0.73. The confusion matrix of this DAE is shown in fig 6.

|  |
| --- |
| PCA of Encoded Test Set (Component 2) |
|  |
| PCA of Encoded Test Set (Component 3) |
|  |
| Plot of Encoded Test Set (Bottleneck Size 2 AE Result) |
|  |

Then we plot the Kaplan-Meier Survival Curve (As shown in fig 7) to do the survival anysis and the log rank test to confirm […]. The result of log rank test is p = 0.0021 2.d.f, showing that the result is […]

*The reason of no clear improvement when move to spare AEs could because SAE is design to solve […] problem, but for the RPPA data, the protein expression is 468 with 324 cases, which may not fit for a SAE.*

The whole working flow of AERO-HNSCC does not require any outside input other than the multi-omics dataset, therefore have the potential to extend to wider ranges of application, it does not require knowledge about specific gene or protein, instead, may be helpful to discovery more OS related gene or protein as shown by the “TopGene” method contributed by *Mondol et al. Referencing Tan et al.’s work, we already see supervised AE can perform well on pan-cancer dataset, given the identical nature of AE, i.e., […]* AERO-HNSCC is highly likely to be able to extend to other cancers or even becoming a universal method for all cancers, this all require future work for validation. For now the single type of cancer still have a limited cases, when extending to multiple cancers, we can expecting a much larger dataset, which is benefit for AE to learn more accurate representation of the input data, we can expect a even better performance.

Another thing could do is extend to more multi-omics data. For now AERO-HNSCC only use the clinical and protein expression data, *Tan et al.'s work provide a good start point, although they assuming each omics data contains equal amount of information and should be encoded to the same dimension, which is not verified by their result, the idea of develop separate AE for each omics and do classification on the combination of all encoded omics is worthtiral.*

Although focus on the different cancer at this stage so that the performance cannot be directly compared, some denoising technique could be integrated into AERO-HNSCC and expecting an improvement. In the case of HNSCC we happen to have a balanced sample of dead and alived, but when extend to more other cancers, sampling techniques such as the *RN-SMOTE presented by Arafa et al. would be helpful.*

The initial expectations of this research is to development of tumour micro-environment transcriptomic hypoxia signatures using AEs, however, with the literature review going deeper one important observation is obtained: No directly related works focus on using AE to develop HS, which is for a good reason. If boarder the range, develop HS using other machine learning techinques like LASSO or LASSO Cox Regression is indeed a trend in this area, another topic AE is more offenly used is to directyly capture the OS signal from the entiral multi-omics data. Considering predict prognosis is also a important purpose for biomarkers like HS development, this two erea is condiered as relavant and the following researches is conducted with the AE path, both to meet the AE foucs requirement of the initial expectations but also because it is indeed a path.

Going down with the path of AE, we soon come to another crossroad. The heterogeneity of mulit-omics data requires completely different processing of data and different architecture of AE. In this case, in the middle stage of this research project, the group is divided to two subgroups. The author generally focus on the protein data and Ming focus on the RNA-Seq data.

For the result achieved by the protein side, although the process was tortuous and challenging, the final result was satisfactory after hard work and the guidance of Professor Delmiro. Through the metrics, including the reconstruction loss of AEs, the precision of classifiers and the p value of the log-rank test, we confirm the validation of protein expression in OS prediction, and the efficiency of encoded protein expression in classification task. satisfiable final result of AERO- HNSCC. Through the benchmark comparison, the result is compenhensively compared with three most relavant and state-of-art works, all shown comparable or better performance.

However, if boarder the range of benchmark comparison and compare our work with less relevant but result is better’s work, we do see some space for improvement. Some SIC developed method achieved comparable or even higher precision in 3 classes classification. Some works applied denoising AEs and *Over Sampling Technique to dealing with the unblanced sample issues. When exrend our work to other types of cancers or pan-cancer, these technique may be useful to further improvement the performance of* AERO- HNSCC.

Building on the findings of Tan et al., we have validated the efficacy of AE in analysing pan-cancer datasets. Given the inherent capabilities of AEs, there is a promising avenue for AERO-HNSCC to be applicable across a range of cancer types, potentially evolving into a universal tool for cancer analysis. However, this hypothesis requires thorough future investigations for comprehensive validation. Currently, our focus on a single cancer type limits the diversity of cases. Expanding this scope to include multiple cancer types could significantly increase the dataset size, thereby enhancing the AE’s ability to learn more accurate data representations and potentially improving overall performance.

The performance of some SIC works is messured by the AUC of the ROC of univariate analysis, which is not fully equivant measurement to the classifier precision. -> need to establish a proven equvalant conversion between AUC and precison, or need to conduct univariate analysis for our result for better comparision, or a better metrics to measure the performance of OS prediction.

Among all binary classification works, the binary classes do not have a standard that all research can follow, and lead to a inconsistent of classification target: in our case we set alive/dead after 10 years and dead in 10 years as two groups, but this important index is not follow or reported in all papers, which may cause bias when comparing performance. -> need a standard classification regulation for pan-cancer or for every individual cancers condiering the life expects different for different cancers.

Looking back to the initial expectation of develop biomarkers such as HS, it is important to notice that biomarkers have significant progronisis value and not only limited to OS prediction, but also predict the the effect of a treatment such as radiotherapy or a certain medicine. Some of these tasks requires a good understanding of its target biomarker. This make the identification of biomarkers important again and in this case SIC method cannot be substitute by AE-based method for its directy identification of biomarkers. However, as shown by Mondol et al., we do see the potential of AE-based method could be useful in this task as well. By anylisis which preotein, gene or other multi-omics feature contribute more to the bottoleneck, we can in this way use AE discovery curtail and novel biomarkers. -> This is only reported in others work not included in AERO- HNSCC, may be useful to include in the further researches.

For the gene side of the AERO- HNSCC, result is not equally safisfierable. XGboost, SVM and other classiferes are tried with multiple AE structures, none of then shows a significant better performance comparing with the raw RNA-Seq data or the relevant works. One thing must be noted that for the AE architecture design section, RNA-Seq require a much more complicated design and much more computation resources to train due to is higher dimensionality. To achieve the simialry proportion of hidder layer size as proetien, i.e., a maximum 16384 node dense layer require more than 30 GB VRAM, this limitation prevent me from design deeper and wider AE for RNA-Seq, and may cause the current AE cannot fully capture the OS signal from the nosie RNA data.

On the other hand, if we continue form the signatures already developed, the dimensionality becomes too low to apply AEs. Encoding the signature that contains 64 genes, which are already verified to be relevant to hypoxia or immune to even lower dimension, we may necessarily lost some important information in this process which is curtail for OS prediction, this may be the cuase that on the signarue subset RNA-Seq data the classifier is still not safisfiable. -> In this case we do not seek method to do dimensionality reduction since the dimensionality of signauture is acceptable already, we may need to discovery other machine learning technique to predict OS with these features.

Noticely one important improvement could be made for the RNA side is the data processing. Unlike a comprehensive cleaning and carefully verified imputation made on the protein data, by the end of project, RNA-Seq data still suffering with umbiguous clinical endpoint issues. In the assitent with Xi Chen I disvovered that a patient may conduct multiple times of RNA-Seq with the same submmiter id, in this case simple merged the clinical endpoint with the RNA data is riskly: there is no way to know which result is recored in the raw RNA data files, and therefore cannot determine the clinical endpoint of the patient. -> need a more thoroughly and carefully data explaoration and proccsing.

The overall collaboration of the team is safifable, especially for the first a few weeks. But soon we meet a unactivate and low contribution issue for some members of the team. This could because of the unblanced time arrangement between the group research project and other modules, or some other reasons. This problem shows on the weekly meeting that only a subset of team member are active on the meeting and synchornize the progress and new findings and disscuss the difficulties identified last week. This leads to the most significant aspect that need to be improved, a better workload distribution and participation. It is a group project, everybody are expected to have about equvlant contribution to the project. To improve, […]

:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Strengths | Weaknesses | Key Contributions |
| Jucheng Hu | Technical Proficiency: Demonstrated deep technical knowledge, especially in protein research.  Independence and Initiative: Proactively managed challenges and advanced project goals.  Problem-Solving: Innovatively addressed and solved complex research problems. | Leadership and Motivation: Failed to sufficiently motivate team members or encourage their contributions.  Workload Distribution: Disproportionately handled the bulk of work, limiting others’ involvement.  Collaboration: Limited delegation and collaboration reduced potential team achievements. | Full Protein Side of AERO- HNSCC |
| Xi Chen | Collaborative Spirit: Excellently fosters teamwork, ensuring cohesive effort and shared objectives.  Inquisitiveness: Actively seeks knowledge through questions, enhancing team understanding and problem-solving.  Timely Communication: Consistently synchronizes progress, ensuring the team stays informed and aligned. | Lack of contribution and participation in the early stage of research | Paper writing and RNA result analysis |
| Tianxiang Xiong | Consistent Contribution: Provides steady, reliable input that supports the team's ongoing efforts.  Versatility: Adapts to various tasks as needed, showing a broad range of skills. | Lack of contribution and participation in the early stage of research | Paper writing and RNA result analysis |
| Ming Wang | Consistent Contribution: Provides steady, reliable input that supports the team's ongoing efforts.  Versatility: Adapts to various tasks as needed, showing a broad range of skills. | Absent in the last stage of RNA development and paper writing | RNA data proccing |
| Eric Shi | Did not participant, not known | Did not participant, not known | NaN |