 doi:10.20944/preprints202210.0351.v1

Communication

Machine Learning Heuristics on Gingivobuccal Cancer Gene Datasets Reveals Key Candidate Attributes for Prognosis

Tanvi Singh*[1], Girik Malik*[1,2,8], Saloni Someshwar*[1], Hien Thi Thu Le[3], Rathnagiri Polavarapu[4], Laxmi N Chavali[1], Jayaraman Valadi[1,5], VS Sundararajan[1], Nidheesh M[7], PB Kavi Kishor[6] and Prashanth Suravajhala[1,7]

1. Bioclues.org, Hyderabad, India
2. Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA
3. Molecular Signaling Lab, Faculty of Medicine & Health Technology, Tampere University, Finland
4. Amity University Rajasthan, Jaipur, India
5. Department of Computer Science, FLAME University, Pune, MH, India
6. Department of Biotechnology, Vignan's Foundation for Science, Technology & Research, Vadlamudi, Guntur, India
7. Amrita School of Biotechnology, Amrita Vishwavidyapeetham, Clappana PO 690525, Kerala, India
8. Labrynthe Pvt. Ltd., New Delhi, India

*Equal contributing authors.

* Correspondence: prash@bioclues.org and pbkavi@yahoo.com

Abstract: Delayed cancer detection is one of the common causes of poor prognosis in case of many cancers including the cancers of the oral cavity. Despite improvement and development of new and efficient gene therapy treatments, very little has been done to algorithmically assess the impedance of these carcinomas. In this work, we attempt to annotate viable attributes in oral cancer gene datasets for identification of gingivobuccal cancer (GBC). We further apply supervised and unsupervised machine learning methods to the gene datasets revealing key candidate attributes for GBC prognosis. Our work highlights the importance of automated identification of key genes responsible for GBC that could perhaps be easily replicated to other forms of oral cancer detection.

Keywords: oral cancer; machine learning; gene prioritization; genomic datasets; data mining

Introduction

Oral cavity cancer (OCC) is the tenth most common malignant tumor in the world and the third most common in southeast Asia. The common subsite recorded in OCC in third world countries, especially in Indian communities is gingivobuccal cancer (GBC) constituting about 40% of all cases, whereas the cases diagnosed in the western world are about 10% (1). They are usually associated with delayed clinical detection, poor prognosis, absence of specific biomarkers for the disease, and expensive therapeutic alternatives (2). The GBC comprises buccal mucosa, gingivobuccal sulcus, alveolus, and retromolar area cancers and is commonly seen in younger patients. While certain precancerous conditions and lesions such as submucous fibrosis, leukoplakia, and erythroplakia are known causes, the dietary deficiencies such as iron, Vitamins A, C, and E are associated with oral cancers. The processes such as segregation of chromosomes, genomic copy number, loss of heterozygosity, telomere stabilities, regulations of cell-cycle checkpoints, DNA damage repairs, and defects in Notch signaling pathways are involved in causing oral cancer (3). Malignant odontogenic tumors emanate *de novo* within jawbones, from epithelium contained within cyst linings, or from malignant transformation of benign odontogenic tumors. The lesions most commonly are the primary intraosseous carcinomas and include the mucoepidermoid carcinoma arising within the bone, and the ameloblastic carcinoma (4). The WHO classification of odontogenic carcinoma dissects malignant ameloblastoma from primary intraosseous carcinoma (5). As diagnosis is entrenched by a biopsy of the jaw lesion, the definitive analysis prospective is of a usually poor outcome. Early signs

 doi:10.20944/preprints202210.0351.v1

and symptoms include soreness or pain in jaws which could extend through chewing/swallowing followed by loosening of teeth and bleeding from mouth. While a good examination is heralded by visualization in the buccal mucosa, the current high end transoral robotic surgeries (TORS) besides vaccines have been in use (6).

Over the last decade, several treatments have been put in use with consistent use of effective gene therapies. New discoveries about how changes in the DNA of cells in the oral cavity and oropharynx cause these cells to become cancerous are being applied to experimental treatments intended to reverse these changes. For example, clinical trials are testing whether it is possible to replace abnormal tumor suppressor genes (such as the p53 gene) of oral cancer cells with a normal copy, to restore normal growth control (7). Machine learning is a computational method that improves performance to make accurate predictions when data analysis and statistical methods do not have enough information about the underlying distribution of data (8). Furthermore from our previous experience, machine learning algorithms have been applied to various fields in genomics (9), healthcare (10), computer vision (Malik et al. 2021) etc. As the applications of these methods have assisted precision medicine scale, this would eventually bridge the gaps in oral squamous cell carcinoma (12). Ahmed et al have earlier investigated these methods from the AI dental imaging perspective. The metadata constituting characteristics, study and control groups were extracted for feature selection paradigms which resulted in understanding the implications of the OSCC. Nevertheless, AI could predict failures to assess the clinical performance in such carcinomas (13). Through the use of statistical methods, the variables (weights) in the algorithm undergo systematic updates representing the distribution of the training data during the training phase. The test phase presents unique unseen data to the same algorithm weights and makes a classification/prediction for this new data point. As these algorithms can help uncover key insights within data mining projects, subsequent decision-making drives can ideally impact key growth metrics. In the present work, we employed a mixture of supervised and unsupervised algorithms and attempted to understand key attributes for prognosis of oral cancer. While supervised methods are much simpler and straightforward to use for our study, we wanted to briefly touch upon the usefulness of unsupervised methods for motivating further research with this combination of data. A detailed gist of results employing Support Vector Machine (SVM), Naïve Bayes, Decision trees, Multi Layer Perceptron, Logistic Regression and K Means (unsupervised) are discussed.

Materials and Methods

Datasets and transformation

We used datasets for four genes related to oral cancer: *PIK3CA*, *KRAS*, *TP53* and *Gingival*. The dataset has the following five features: (i) name, (ii) gene(s), (iii) protein change, (iv) condition(s), clinical significance (Last reviewed). TP53 and Gingival have an additional Review Status feature. Number of samples vary for each dataset: PIK3CA has 544 instances, KRAS has 330 instances, TP53 has 2186 instances and Gingival has 2107 instances (Table S1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 October 2022 doi:10.20944/preprints202210.0351.v1

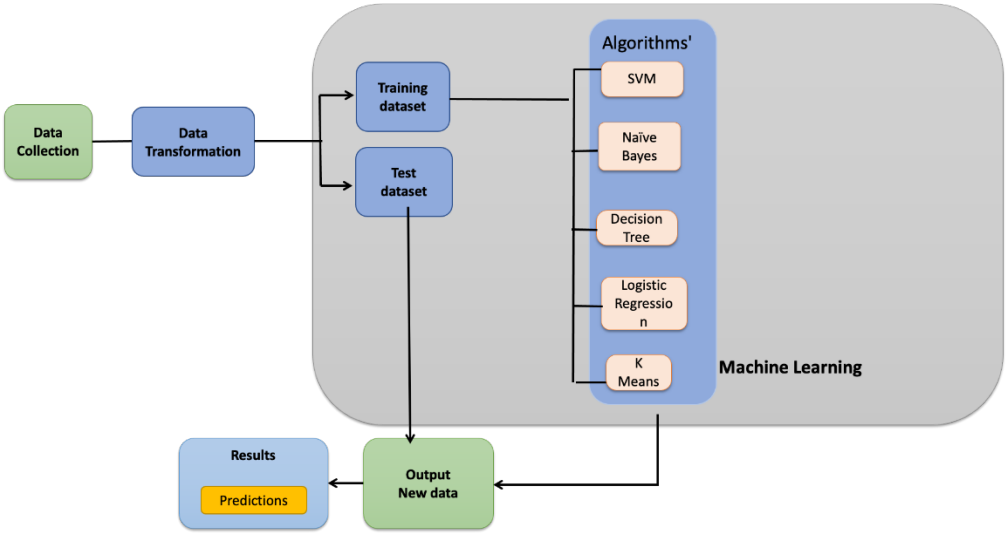

Figure 1: Machine Learning pipeline used for our analysis.

We transformed alphanumeric features into categorical features for application of the following Machine Learning Algorithms (as given below in section Classifier design and training). The first instance of data values of protein change, condition(s), clinical significance (last reviewed) and review status was used. Then data values of features such as gene(s), protein change, condition(s), clinical significance (last reviewed) and review status were converted into numeric keys using Preprocessing and Transformation classes in scikit-learn. Binary and numeric weightages were assigned to each of the features including protein change, condition(s), clinical significance (last reviewed) and review status to evaluate the performance based on data annotations.

Experiments

We performed four experiments for PIK3CA and KRAS datasets, and six experiments for TP53 and Gingival using different combinations of features. The following six experiments separately used one of the following four features: (i) all the features in a dataset, (ii) only binary features, (iii) only non-binary features, (iv) all features except review status (for datasets (TP53, Gingival) that contains review status as a feature) (v) only non-binary features with no review status (for datasets (TP53, Gingival) that contains review status as a feature), (vi) only binary features with no review status (for datasets (TP53, Gingival) that contains review status as a feature)

Classifier design and training

We used six major classifiers to train and test the model: (i) Support Vector Machine (ii) Naïve Bayes (iii) Decision trees (iv) Perceptron (v) Logistic Regression and (vi) K Means (unsupervised). We randomly split the dataset to use 80% for training and 20% for testing. We used off-the-shelf algorithms implemented in scikit-learn for these experiments and used other libraries like numpy, pandas, and matplotlib available in Python. While unsupervised algorithms are hard to implement on such data, we used only K Means for a flavor of unsupervised learning. Further analyses with algorithms like K Medoids, PCA, etc. are left for future work.

Performance evaluation

Evaluating the performance of learning algorithms is a central aspect of machine learning. We used an 80-20 train-test split to test the performance of the predictive and classification models. To mitigate the overfitting problem, the following measures were used to evaluate the performance six classifiers based on accuracy which is defined as the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 October 2022 doi:10.20944/preprints202210.0351.v1

percentage of correct predictions for the test data. It can be calculated by dividing the number of correct predictions by the number of total predictions. The measure is defined as follows:

Accuracy = (TP + TN) / (TP + FN + FP + TN)

where TP: True Positives (positive samples classified correctly as positive), TN: True Negatives (negative samples classified correctly as negative), FP: False Positives (negative samples predicted wrongly as positive) and FN: False Negatives (positive samples predicted wrongly as negative). The precision and recall were achieved with inherent accuracy.

Results and Discussion

PIK3CA among the select genes with highest accuracy

One of the interesting findings that we attempted in our study was to identify those gene datasets that are significantly enriched from machine learning heuristics. We observe that there is a significant amount of attribute fitting with instances taken up from all datasets. While all instances were used and compared across all the algorithms, to further gain insights into this, the accuracies were tabulated accordingly (Figure 2; supplementary table 1). For PIK3CA, experiment (i) accuracy varies between 78% (decision tree) and 48% (Naïve Bayes). For experiment (ii) accuracy varies between 67% (MLP) and 41% (Naïve Bayes). For experiment (iii) accuracy varies between 77% (decision tree) and 44% (Naïve Bayes). On the other hand, for KRAS, experiment (i) accuracy varies between 62% (decision tree) and 27% (K Means). For experiment (ii) accuracy varies between 62% (decision tree) and 17% (Naïve Bayes). For experiment (iii) accuracy varies between 53% (decision tree) and 18% (Naïve Bayes). Whereas TP53 showed variable changes, for experiment (i) accuracy varies between 61% (MLP) and 35% (K Means). For experiment (ii) accuracy varies between 56% (SVM, MLP and decision tree) and 35% (K Means). For experiment (iii) accuracy varies between 55% (MLP) and 8% (Naïve Bayes). For experiment (iv) accuracy varies between 57% (MLP) and 34% (K Means). and for experiment (v) accuracy varies between 50% (decision tree) and 21 (K Means). For experiment (vi) accuracy varies between 51% (decision tree, logistic regression, MLP, SVM) and 35% (K Means). For gingival datasets, experiment (i) accuracy varies between 63% (MLP) and 29% (K Means), for experiment (ii) accuracy varies between 49% (MLP) and 29% (K Means), for experiment (iii) accuracy varies between 63% (decision tree) and 29% (K Means), for experiment (iv) accuracy varies between 54% (MLP) and 29% (K Means), for experiment (v) accuracy varies between 52% (MLP) and 29% (K Means) and for experiment (vi) accuracy varies between 40% (decision tree, logistic regression, MLP, SVM) and 30% (K Means clustering). From above results it is evident that only experiment (i) is shown to have highest accuracy when compared with other experiments from (ii) to (vi) (Table 1)

Table 1: ML Accuracies for each Oral Cancer Genes.

		ML Algorithms (Accuracies %)					
		SVM	MLP	Logistic regression	Naïve Byes	Decision Tree	K-means Unsupervised
Genes	PIK3CA	71	66	56	48	78	48
	KRAS	41	55	39	17	62	27
	TP53	56	63	54	48	58	29
	Gingival	53	61	42	54	53	35

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 October 2022 doi:10.20944/preprints202210.0351.v1

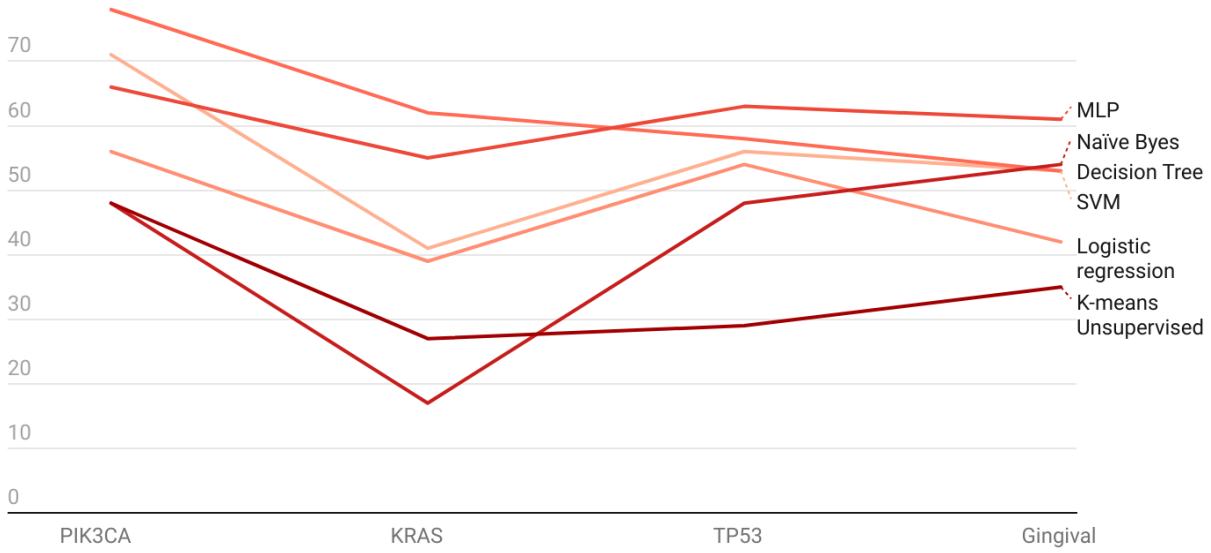

Figure 2: Machine Learning accuracies for vivid datasets exploited using various experiments

What we aimed to achieve from our pilot study is to employ gene selection and ask whether or not the lesser known changes in attributes can by choice be ignored for further prognosis. In other words, in nature, are there any genes that are repetitively expressed with inherent changes, attempted in our machine learning heuristics (15). The virtual experiments on ML heuristics which we employed sets a base for oral cancer prognosis, however there is a dearth of well annotated or informative attributes which is a major limitation of our work. Theoretically, with more instances and genes segregated from the attributes, we could have gotten a better performance and overcome the overfitting problem albeit the fact that our finding the relevant four genes augments the hypothesis that it may not always be true. Our experiments and framework can further be extended for revealing the effects of key attributes from genetic data, and be applied to predict outcomes like the chances of survival, recurrence, etc. On the other hand, some work has been seen around survival risk stratification (16), and survival prediction (17) using similar machine learning based methods. Majority of these works have patient datasets collected for several years even as these yield *bona fide* results, they could be prone to biases. We found the application of Principal Component Analysis (PCA) and other techniques for data reduction to be prevalent in multiple studies.

Initially, we ran the experiments with the same data splits and the same machine learning algorithms using the java-based package Weka (18,19). While we found the results to be clearly overfitting to our data, we speculate that Weka assigns every non-numeric instance to be a unique key and processes them individually. For example, when A-B was arranged as B-A in the dataset (without ordering-sensitive features), Weka is unable to break them and considers them as two keys instead of one. A clear limitation for this approach is indicative of certain data types, as it also relies heavily on data annotation. Having data annotated (manually and programatically) to account for such orderings, we find that our models do not overfit and perform better which could be the plausible reason why many annotated cancer datasets have. This is also in agreement with the fact that the scarcity of publicly available image datasets may impede early patho-significant diagnoses for cancers taking the machine learning paradigm (20). On the other hand, to overcome the overfitting and failed model as we postulated, deep learning models could bring great promise for accurate prognosis, if in case the datasets have tumorigenic data, infiltrating lymphocytes and multiclass labeling which can herald predicting disease states (13). Such data could then be divided into risk groups and then differentiate the data from good to poor prognosis. Having said this, deep learning clubbed with precise

detection may then be used to identify oral cancer datasets albeit the fact that there must be high-end computability to identify multidimensional datasets.

Given the success of multimodal algorithms (21), we believe our analysis can be further strengthened by using microscopic images of cells from the buccal cavity alongside annotated genetic data. Using electron microscopy and image segmentation algorithms, it is now possible to segment the image upto cellular level, precisely pinpointing the areas of carcinoma. Such precise positions can help prevent the pitfalls of annotation errors, making our analysis more robust. We speculate that such analysis can also aid in the prediction of early onset of cancers (22, 14).

Conclusions

Oral cancer prognosis is one of the burgeoning problems and our work employing machine learning heuristics could lay emphasis in piloting candidate biomarkers. As diagnosis could be better aided for prognosis and theranostics, survival and therapies must be in place and despite strategic improvements in these areas, this is still in infancy. Machine learning and artificial intelligence (AI) aided methods have enhanced early detection in reducing mortality and morbidity. Indefatigably, there are not many metadata based machine learning heuristics assessing the impedance of these carcinomas. In summary, we presented a machine learning based approach to predict the gene dataset which reveals key candidate attributes for GBC prognosis. We have attempted to fill these gaps by performing and labeling classes, accurate identification of viable attributes for such cancers. Furthermore, we found that deterministic methods perform well with limited data, while non-deterministic methods excel in performance with large datasets wherein supervised learning methods perform better than unsupervised methods. Nonetheless, our experiments had more supervised methods than unsupervised ones, which we wanted to establish the use case for such an analysis. We argue that a multitude of unsupervised and semi-supervised methods might be able to better model these data distributions which seldom have accurate annotations. However, this may be due to the lack of machine learning heuristics which could be used as models and *vice versa* for a better modeled framework.

Competing interests: None

Authors' contributions: TB, SS and GM contributed equally to the work and ran the ML heuristics. GM conceptualized the experiments. SS and HT annotated the datasets. GM revised the experiment plan. RP, LN, JV, VS,NM curated the ML heuristics, PS, PBK and VKJ co supervised the work

Funding: None

Acknowledgements: None

Competing interests: None

References

1. Mandlik Dushyant S, Nair Suraj S, Patel Kaustubh D, Gupta Karan, Patel Purvi, Patel Parin, Sharma Nitin, Joshipura Aditya, Patel Mitesh. Squamous cell carcinoma of gingivobuccal complex: Literature, evidences and practice. 2018: ^(1): 18-28.
2. Rivera C. Essentials of oral cancer. *Int J Clin Exp Pathol*. 2015;8(9):11884-11894. Published 2015 Sep 1.
3. Ali J, Sabiha B, Jan HU, Haider SA, Khan AA, Ali SS. Genetic etiology of oral cancer. Oral Oncol. 2017 Jul;70:23-28. doi: 10.1016/j.oraloncology.2017.05.004. Epub 2017 May 17. PMID: 28622887.
4. Clayman L. Malignant Odontogenic Tumors. In: Kufe DW, Pollock RE, Weichselbaum RR, et al., editors. Holland-Frei Cancer Medicine. 6th edition. Hamilton (ON): BC Decker; 2003. Available from: https://www.ncbi.nlm.nih.gov/books/NBK13124/
5. Elzay RP . Classification of primary intraosseous carcinoma. Oral Surg Oral Med Oral Pathol. 1982;54:299–303.
6. Nguyen AT, Luu M, Mallen-St Clair J, et al. Comparison of Survival After Transoral Robotic Surgery vs Nonrobotic Surgery in Patients With Early-Stage Oropharyngeal Squamous Cell Carcinoma. *JAMA Oncol.* 2020;6(10):1555–1562. doi:10.1001/jamaoncol.2020.3172
7. Deshpande AM, Wong DT. Molecular mechanisms of head and neck cancer. Expert Rev Anticancer Ther. 2008 May;8(5):799-809. doi: 10.1586/14737140.8.5.799. PMID: 18471051; PMCID: PMC2709830.

8. Mohri, Mehryar, Afshin Rostamizadeh and Ameet S. Talwalkar. "Foundations of Machine Learning." Adaptive computation and machine learning (2012).
9. Ijaq, J., Malik, G., Kumar, A., Das, P.S., Meena, N., Bethi, N., Sundararajan, V.S. and Suravajhala, P., 2019. A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. *BMC bioinformatics*, *20*(1), pp.1-8.
10. Malik, G., Gulati, IK, 2020. Little Motion, Big Results: Using Motion Magnification to Reveal Subtle Tremors in Infants. *Workshop on Artificial Intelligence for Healthcare in 24th European Conference on Artificial Intelligence*.
11. Malik, G., Linsley, D., Serre, T., Mingolla, E., 2021. The Challenge of Appearance-Free Object Tracking with Feedforward Neural Networks. *CVPR Workshop on Dynamic Neural Networks Meet Computer Vision*.
12. Alabi RO, Almangush A, Elmusrati M, Mäkitie AA. Deep Machine Learning for Oral Cancer: From Precise Diagnosis to Precision Medicine. Front Oral Health. 2022 Jan 11;2:794248. doi: 10.3389/froh.2021.794248. PMID: 35088057; PMCID: PMC8786902.
13. Ahmed N, Abbasi MS, Zuberi F, Qamar W, Halim MSB, Maqsood A, Alam MK. Artificial Intelligence Techniques: Analysis, Application, and Outcome in Dentistry-A Systematic Review. Biomed Res Int. 2021 Jun 22;2021:9751564. doi: 10.1155/2021/9751564. PMID: 34258283; PMCID: PMC8245240.
14. de Guia, J.M., Devaraj, M. and Leung, C.K., 2019, August. DeepGx: deep learning using gene expression for cancer classification. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 913-920).
15. Mahendran N, Durai Raj Vincent PM, Srinivasan K and Chang C-Y (2020) Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues, and Future Research Directions. Front. Genet. 11:603808. doi: 10.3389/fgene.2020.603808
16. Tseng Y, Wang H, Lin T, Lu J, Hsieh C, Liao C. Development of a Machine Learning Model for Survival Risk Stratification of Patients With Advanced Oral Cancer. *JAMA Netw Open.* 2020;3(8):e2011768. doi:10.1001/jamanetworkopen.2020.11768
17. Kim, D.W., Lee, S., Kwon, S. *et al.* Deep learning-based survival prediction of oral cancer patients. *Sci Rep* 9, 6994 (2019). https://doi.org/10.1038/s41598-019-43372-7
18. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
19. I.H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, 2000.
20. Sengupta N, Sarode SC, Sarode GS, Ghone U. Scarcity of publicly available oral cancer image datasets for machine learning research. Oral Oncol. 2022 Mar;126:105737. doi: 10.1016/j.oraloncology.2022.105737. Epub 2022 Feb 2. PMID: 35114612.
21. Nigam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, January). Multimodal deep learning. In *ICML*.
22. Krause, J., Grabsch, H.I., Kloor, M., Jendrusch, M., Echle, A., Buelow, R.D., Boor, P., Luedde, T., Brinker, T.J., Trautwein, C. and Pearson, A.T., 2021. Deep learning detects genetic alterations in cancer histology generated by adversarial networks. *The Journal of pathology*, *254*(1), pp.70-79.