A Stacked Ensemble Learning Approach with Two-Stage Gene Selection for Identifying Biomarkers and Molecular Pathways in Oral Cancer and Periodontal Disease

Mst. Amina Khatun[1], Shreestha Pallabi Hossain[2] and Md.Faruk Hosen[3]*

[1]Department of Computing and Information System, Daffodil International University, Savar, Dhaka 1216, Bangladesh

[2]Department of Software Engineering, Daffodil International University, Savar, Dhaka 1216, Bangladesh

[3]Department of Computer Science and Engineering, Begum Rokeya University, Rangpur - 5404, Bangladesh

***Corresponding Authors: Md. Faruk Hosen (faruk@cse.brur.ac.bd)**

BACKGROUND

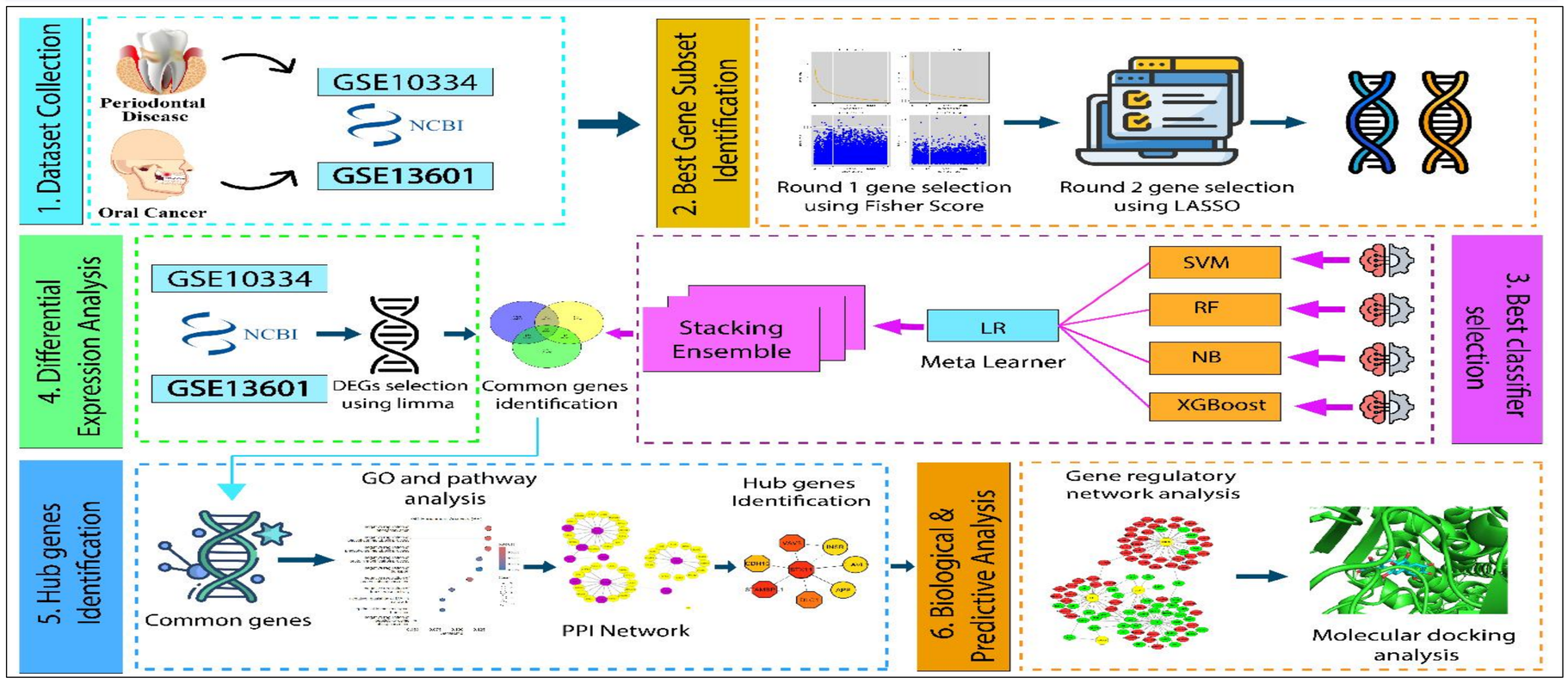
Oral cancer (OC), a prevalent malignancy, is characterized by uncontrolled cellular proliferation in oral tissues, causing significant mortality with a 5-year survival rate of **only ~50%**. Affecting 10.5–12% of the global population, it often results from late-stage diagnosis or co-existing inflammatory conditions like **Periodontal disease (PD)**. Current diagnostic approaches may have limitations such as high dimensionality of genomic data and inability to identify shared molecular biomarkers between OC and PD, especially in patients with complex transcriptomic profiles.

FS–LASSO, a hybrid feature selection framework, is commonly developed in computational studies to identify significant biomarkers and evaluate potential therapeutic targets. Stacking ensemble learning, an advanced machine learning strategy with various analytical strengths, such as multi-classifier integration, improved generalization, reduced model bias and variance, and robust performance on high-dimensional transcriptomic data.

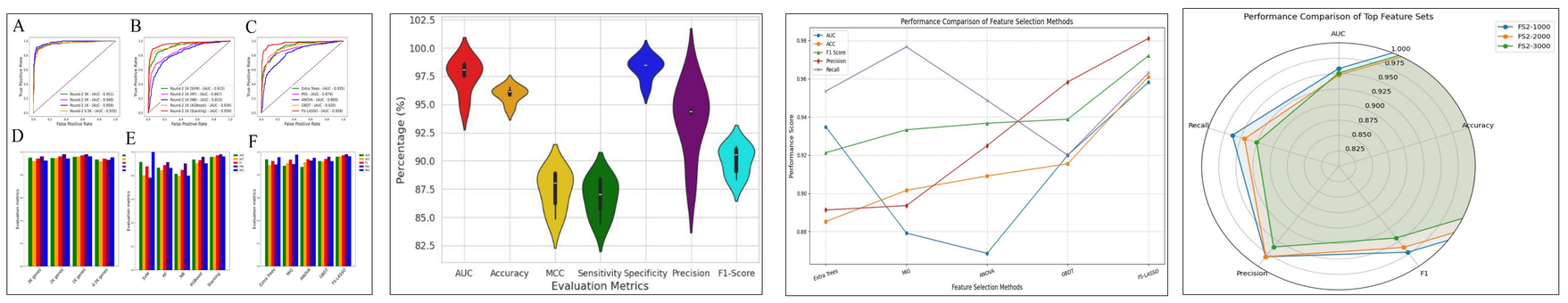
AIMS AND OBJECTIVES

- Develop a hybrid FS–LASSO framework combining Fisher Score and LASSO to extract informative gene subsets.
- Employ a Stacking ensemble model integrating SVM, Random Forest, Naive Bayes, and XGBoost for robust classification.
- Identify common DEGs shared between oral cancer and periodontal disease through bioinformatics analysis.
- Perform GO annotation and KEGG, Reactome, and WikiPathways analysis to uncover molecular mechanisms.
- Conduct molecular docking of hub genes with FDA-approved compounds to propose therapeutic candidates for oral cancer.

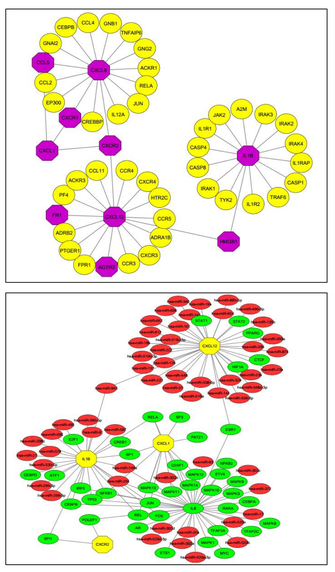
METHODOLOGY



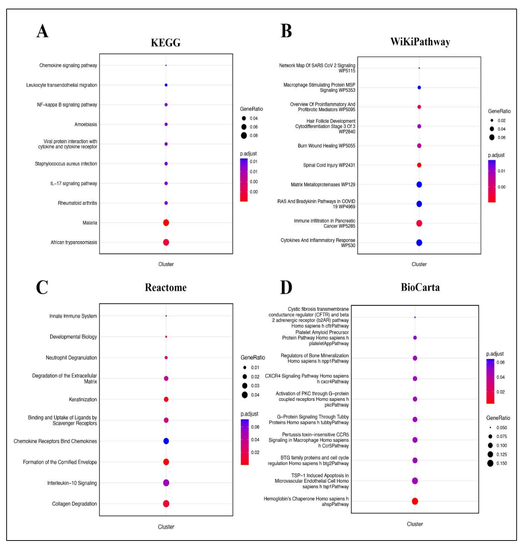
RESULTS



Figure(1): Performance comparison of feature selection techniques and classification models



Figure(2): Gene-regulatory network analysis



Figure(3): Pathway analysis

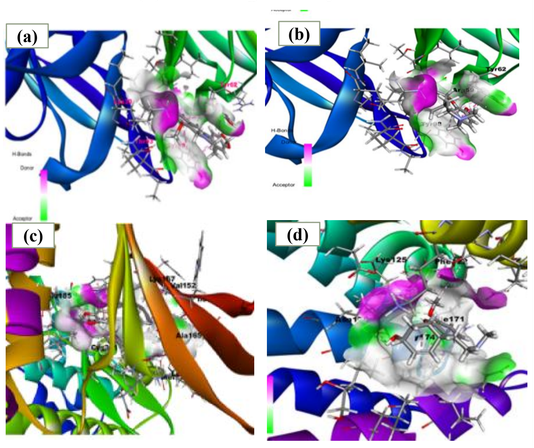
Phytochemicals	Binding Energy (ΔG =kcal/mol) to TSHR	Binding Energy (ΔG=kcal/mol) to D2
Caftaric acid	-6.7	-6.7
6 - Benzylaminopurine	-6.3	-6.5
Oxitriptan	-5.8	-6.2
Thyroxine (control)	-5.3	-6.0

Table(1): Binding affinity scores

Figure(4): Gene expression visualization of OC

Stage	Method	AUC	ACC	F1 Score	PRE	REC
Without FS	SVM	0.8592	73.77%	0.8367	0.7455	0.9535
	RF	0.8482	80.43%	0.8571	0.8710	0.8438
	NB	0.6200	74.03%	0.8333	0.7576	0.9259
	XGBoost	0.8308	82.61%	0.8806	0.8551	0.9077
	Stacking	0.9419	85.92%	0.8936	0.9545	0.8400
FS Round 1	SVM	0.8708	77.05%	0.8600	0.7544	1.0000
	RF	0.8349	81.82%	0.8611	0.9394	0.7949
	NB	0.6389	78.69%	0.8687	0.7679	1.0000
	XGBoost	0.9315	88.52%	0.9195	0.9091	0.9302
	Stacking	0.9521	90.22%	0.9291	0.9516	0.9077
FS Round 2	SVM	0.9148	80.23%	0.8777	0.7821	1.0000
	RF	0.8667	84.62%	0.8889	0.9143	0.8649
	NB	0.8153	79.59%	0.8485	0.9032	0.8000
	XGBoost	0.9362	90.54%	0.9307	0.9592	0.9038
	Stacking	0.9581	96.10%	0.9720	0.9811	0.9630

Table(2): Stacking and Baseline Classifier Comparison Across FS Stages



Figure(5): Molecular docking representations

CONCLUSION

This study provides compelling evidence for the potential of the FS–LASSO framework in identifying shared biomarkers between oral cancer and periodontal disease, achieving **96.10%** classification accuracy and an AUC of 0.9581. Hub genes CXCL8, IL1B, CXCR2, CXCL1, and CXCL12 showed strong binding affinities with FDA-approved drug compounds, warranting further investigation as promising therapeutic targets for oral cancer treatment..

FUTURE PLAN

1. Larger RNA-seq datasets will be incorporated to improve the generalizability of the proposed FS–LASSO framework.
2. Advanced deep learning techniques will be employed to further validate the therapeutic potential of identified hub genes through wet-lab experimental studies.

REFERENCES

1. Chang, T. S., Chang, C. M., Ho, H. C., Su, Y. C., Chen, L. F., & Chou, P. (2013). Impact of young age on the prognosis for oral cancer: a population-based study in Taiwan. PLoS One, 8(9), e75855.
2. Cristaldi, M., Mauceri, R., Di Fede, O., Giuliana, G., Campisi, G., & Panzarella, V. (2019). Salivary biomarkers for oral squamous cell carcinoma diagnosis and follow-up: current status and perspectives. Frontiers in Physiology, 10, 1476.