Exploring the Supervised Learning Models to Automatically Diagnose Colon Cancer Patients based on their SNP Profiles

Ali Reza Ibrahimzada[1], Huzeyfe Ayaz[1], Seyda Demirkol[2,3], Dilara Sönmez[2], Soykan Arıkan[4], Mehmet Tolgahan Hakan[2,5], Özlem Küçükhüseyin[2], Saime Turan Sürmen[2], İlhan Yaylım[2], Ali Cakmak[1] , Mehmet Baysan[1]

[1]Department of Computer Science and Engineering, Istanbul Sehir University, Istanbul, Turkey

[2]Molecular Medicine Department, Aziz Sancar Institute of Experimental Medicine / Istanbul Medical Faculty at Istanbul University, Istanbul, Turkey.

[3]Istanbul Biruni University , Vocational School Of Health Services, Istanbul- TURKEY

[4]General Surgery Clinics, Istanbul Training and Research Hospital, Istanbul, Turkey.

[5] Hitit University, Art and Science Faculty, Department of Biology.Çorum , Turkey.

The goal this analysis is to explore the machine learning-based automatic diagnosis of colorectal patients based on the single nucleotide polymorphisms (SNP). Such a computational approach may be used complementary to other diagnosis tools, such as, biopsy, CT scan, and MRI. Moreover, it may be used as a low-cost screening for colorectal cancers to improve the public health.

Dataset: The dataset includes SNPs observed in particular genomic loci that are located within DNA regions of 11 selected genes, namely, p16540, p16580, mdm2, GAL3, TIM1, trail, pd1duz, pdl1poly, CD28, cd27snp, and CD40. The dataset included 50 healthy individuals (control group) and 65 colon cancer patients. The dataset also includes additional information for patients only, such as, the age of the patient, the stage of the cancer (in terms of two different staging models, namely, T-model and n-model), the location of the tumor, peripheral invasiveness of the tumor, as well as tissue-differentiation around tumor regions.

Methods: We employ several supervised classification algorithms, namely, Logistic Regression (LR), Random Forests (RF), and Support Vector Machine (SVM). Besides, we apply knn-based data imputation to fill the missing genotype values. The scripts are written in Python programming language using the Scikit-learn library. To evaluate different approaches and models, we exploit f1-scores and Area Under Curve (AUC) values in ROC curves.

Conclusion: We make the following observations:

(1) Logistic Regression-based classifier using one-hot encoding for feature representation and knn-based imputation to complete the missing data performs the best among the studied classifiers in terms of both f1-score (88%) and AUC value (0.88).
(2) Knn-based data imputation increases the f1-scores around 18% (from 70% to 88%).
(3) Based on the high accuracy of the constructed logistic regression models, the studied 11 genes may be considered as a gene panel candidate for the diagnosis of colon cancer.
(4) Based on the ANOVA analysis, the following genotypes are the statistically significant (i.e., p-val < 0.05) discriminating features between the control group and colon cancer patients. This shows that the colon cancer is associated with multiple genes in complex interactions.

	f-values	p-values
CD40_C/C	37.50834	1.36e-08
CD28_C/C	18.61001	3.44981e-05
CD40_C/T	14.33445	0.0002467735
pd1duz_C/T	10.51366	0.0015578771
trail_C/T	7.28887	0.0080047526
trail_C/C	7.28887	0.0080047526
CD28_T/T	6.78186	0.0104460442
cd27snp_T/T	6.15964	0.0145409241
pdl1poly_A/C	5.70264	0.0185977448
pdl1poly_A/A	5.64805	0.0191561912
GAL3_C/C	4.85052	0.0296682817