RESULTS

Table 1: Evaluation Criteria for Random Forest Classifier for each of the three datasets

Dataset	Kappa	Precision	Accuracy	Specificity	Sensitivity	Overall Error
Leukaemia (GSE9476)	90.22	96.00	95.31	97.37	92.31	4.69
Colon Cancer (GSE44861)	74.76	86.21	87.39	85.45	89.29	12.61
Prostate Cancer (GSE71783)	40.0	71.43	70.0	73.33	66.67	30.0

Gene Function Analysis for Leukemia Cancer Dataset (GSE9476)

Table 2: The Smallest Set of Genes Selected From Leukemia Cancer Data by the Random Forest Method

Gene Symbol	Function
ALDH1A1	Aldehyde Dehydrogenase 1 Family Member A1. The protein encoded by this gene is involved in alcohol and retinol metabolism. [1] ALDH1A1) is highly expressed in CD34$^+$ hematopoietic stem cells and is involved in important functions that maintain the stem cell compartment. Loss of ALDH1A1 has been found to cause Acute Myeloid Leukemia (AML) in murine model. [2] Yang et. al have reported that Alcohol Dehydrogenases are potential molecular markers of AML and may also be involved in drug resistance in the same. [3]
BAG4	BAG cochaperone 4 The protein encoded by BAG4 is an anti-apoptotic protein and it interacts with many cell apoptosis and growth-related proteins. [4] BAG4/silencer of death domains (SODD) plays an important role in apoptosis and chemotherapy sensitivity. Inhibition of its activity is found to reverse multidrug resistance in AML. [5]
GPX1	Glutathione Peroxidase 1 The protein encoded by the gene GPX1 protects cells from oxidative damage. [6] It also protects hemoglobin in erythrocytes from oxidative breakdown. [7] Polymorphisms in GPX1 may influence the risk of developing AML.[8] Wei et al. have observed that high levels of GPX1 is associated with poorer prognosis of overall survival in AML patients. [9]
JAG1	Jagged Canonical Notch Ligand 1 The protein encoded by JAG1 is involved in signalling processes, hematopoiesis and in early and late stages of mammalian cardiovascular development. [10] Abnormalities in Notch1 and JAG1 proteins is associated with abnormal proliferation of AML cells. [11] JAG1 levels could be a potential prognostic factor that could predict survival in AML[12]
PLXNC1	Plexin C1 The gene encodes proteins of Plexin family which are transmembrane receptors for semaphorins. Semaphorins are proteins that regulate immune responses, cell motility and axon guidance. [13] Stirewalt et al. have observed that AML cells display decreased PLXNC1 transcript levels when compared to normal hematopoietic cells. [14]

References:

1. https://www.ncbi.nlm.nih.gov/gene/216

2. Maura Gasparetto, Shanshan Pei, Mohammad Minhajuddin, Daniel A. Pollyea, Vasilis Vasiliou, R. Keith Humphries, Craig T. Jordan, Clayton A. Smith; Aldehyde Dehydrogenases Play a Role in Acute Myeloid Leukemia and Have Prognostic and Therapeutic Significance. Blood 2014; 124 (21): 2238.

https://ashpublications.org/blood/article/124/21/2238/90088/Aldehyde-Dehydrogenases-Play-a-Role-in-Acute

3. Xiangchou Yang, Rongxin Yao, Hong Wang, "Update of ALDH as a Potential Biomarker and Therapeutic Target for AML", *BioMed Research International*, vol. 2018, Article ID 9192104, 5 pages, 2018. https://www.hindawi.com/journals/bmri/2018/9192104/

4. https://www.ncbi.nlm.nih.gov/gene/9530

5. Deng L, Jiang L, Lin XH, et al. The PI3K/mTOR dual inhibitor BEZ235 suppresses proliferation and migration and reverses multidrug resistance in acute myeloid leukemia. *Acta Pharmacol Sin*. 2017;38(3):382-391. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5342661/

6. https://www.ncbi.nlm.nih.gov/gene/2876

7. https://www.uniprot.org/uniprot/P07203

8. Claudia Bănescu, Mihaela Iancu, Adrian P. Trifa, Marcela Cândea, Erzsebet Benedek Lazar, Valeriu G. Moldovan, Carmen Duicu, Florin Tripon, Andrei Crauciuc, Minodora Dobreanu, "From Six Gene Polymorphisms of the Antioxidant System, Only GPX Pro198Leu and GSTP1 Ile105Val Modulate the Risk of Acute Myeloid Leukemia", *Oxidative Medicine and Cellular Longevity*, vol. 2016. https://www.hindawi.com/journals/omcl/2016/2536705/

9. Wei J, Xie Q, Liu X, et al. Identification the prognostic value of glutathione peroxidases expression levels in acute myeloid leukemia. *Ann Transl Med*. 2020;8(11):678. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7327321/

10. https://www.uniprot.org/uniprot/P78504

11. Tohda S, Nara N. Expression of Notch1 and Jagged1 proteins in acute myeloid leukemia cells. Leuk Lymphoma. 2001 Jul;42(3):467-72. https://pubmed.ncbi.nlm.nih.gov/11699411/

12. Czemerska M, Pluta A, Szmigielska-Kaplon A, Wawrzyniak E, Cebula-Obrzut B, Medra A, Smolewski P, Robak T, Wierzbowska A. Jagged-1: a new promising factor associated with favorable prognosis in patients with acute myeloid leukemia. Leuk Lymphoma. 2015 Feb;56(2):401-6. https://pubmed.ncbi.nlm.nih.gov/24844362/

13. https://www.ncbi.nlm.nih.gov/gene/10154

14. Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W, Pogosova-Agadjanyan EL, Engel JH, Cronk MR, Dorcy KS, McQuary AR, Hockenbery D, Wood B, Heimfeld S, Radich JP. Identification of genes with abnormal expression changes in acute myeloid leukemia. Genes Chromosomes Cancer. 2008 Jan;47(1):8-20. https://pubmed.ncbi.nlm.nih.gov/17910043/

Gene Function Analysis for Colon Cancer Dataset (GSE44861)

Table 3: The Smallest Set of Genes Selected From Colon Cancer Data by the Random Forest Method

Gene Symbol	Function
CA1	Carbonic Anhydrase 1 Carbonic anhydrases are a large family of zinc metalloenzymes that catalyze the reversible hydration of carbon dioxide. They also participate in many biological processes such as respiration, calcification, acid-base balance, bone resorption, and the formation of aqueous humor, cerebrospinal fluid, saliva and gastric acid. [1] CA1 is expressed in normal colonic mucosa, but its reduced expression correlates with vascular invasion and poor prognosis of colorectal cancers. [2] Recent studies also show a down-regulation of CA1 expression in colorectal cancer cells. [3]
CA7	Carbonic Anhydrase 7 Decreased expression of CA7 is associated with disease progression and predicts poor prognosis in colorectal cancer, especially in patients with early stage tumors. [4]
DIEXF	Digestive-Organ Expansion Factor Homolog It regulates the p53 pathway to control the expansion growth of digestive organs. [5] Hypomethylation of Alu repeats in various cancer types may contribute to alteration of DIEXF in cancer cells and make it a potential tumor biomarker. [6]
GUCA2A	Guanylate Cyclase Activator 2A
GUCA2B	Guanylate Cyclase Activator 2B Protein encoded by GUCA2A is an endogenous activator of intestinal guanylate cyclase.[7] GUCA2B encodes a preproprotein that is proteolytically processed to generate multiple protein products, including uroguanylin. It regulates salt and water homeostasis in the intestine and kidneys. [8] Guanylyl cyclase C is a transmembrane receptor expressed on intestinal epithelial cells, that is involved in intestinal homeostatic mechanisms. These effects are mediated by the endogenous hormones guanylin (GUCA2A) and uroguanylin (GUCA2B), which bind and activate GUCY2C to regulate proliferation, metabolism and barrier function in intestine. Research shows that GUCY2C silencing increases colon cancer susceptibility in animals and humans. [9]

Table 3: The Smallest Set of Genes Selected From Colon Cancer Data by the Random Forest Method (Continued)

Gene Symbol	Function
IGH	Immunoglobulin Heavy Locus Immunoglobulins recognize foreign antigens and initiate immune responses such as phagocytosis and the complement system. Each immunoglobulin molecule consists of two identical heavy chains and two identical light chains. IGH region represents the germline organization of the heavy chain locus. [10] Geng et al. have observed that 5 IgH classes are expressed in both colon cancer cells as well as normal colon epithelial cells but those expressed in cancer cells have unique hypermutation hot points. Their findings suggest that Ig Heavy chains could be used as a novel target in immune therapy for colon cancer. [11]
SPIB	Spi-B Transcription Factor The protein encoded by this gene is a transcriptional activator that binds to the PU-box and acts as a lymphoid-specific enhancer. It increases production of natural interferon (IFN)-producing cells that produce large amounts of interferon and block viral replication. [12] SPIB expression is down-regulated in colon cancer cells. It functions as a tumor suppressor and may increase sensitivity of colon cancer cells to chemotherapy. [13]

References:

1. https://www.ncbi.nlm.nih.gov/gene/759

2. https://www.gastrojournal.org/article/0016-5085(93)90900-W/pdf

3. Osamu Kitahara, Yoichi Furukawa, Toshihiro Tanaka, Chikashi Kihara, Kenji Ono, Renpei Yanagawa, Marcelo E. Nita, Toshihisa Takagi, Yusuke Nakamura and Tatsuhiko Tsunoda. Alterations of Gene Expression during Colorectal Carcinogenesis Revealed by cDNA Microarrays after Laser-Capture Microdissection of Tumor Tissues and Normal Epithelia.
https://cancerres.aacrjournals.org/content/61/9/3544

4. Yang, GZ., Hu, L., Cai, J. *et al.* Prognostic value of carbonic anhydrase VII expression in colorectal carcinoma. *BMC Cancer* 15**,** 209 (2015).
https://bmccancer.biomedcentral.com/articles/10.1186/s12885-015-1216-y

5. https://www.uniprot.org/uniprot/Q68CQ4

6. Martín B, Pappa S, Díez-Villanueva A, et al. Tissue and cancer-specific expression of DIEXF is epigenetically mediated by an Alu repeat. Epigenetics. 2020 Jun - Jul;15(6-7):765-779.
https://europepmc.org/article/pmc/pmc7574385

7. https://www.uniprot.org/uniprot/Q02747

8. https://www.ncbi.nlm.nih.gov/gene/2981

9. Pattison AM, Merlino DJ, Blomain ES, Waldman SA. Guanylyl cyclase C signaling axis and colon cancer prevention. *World J Gastroenterol*. 2016;22(36):8070-8077.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5037076/

10. https://www.ncbi.nlm.nih.gov/gene/3492

11. Geng ZH, Ye CX, Huang Y, et al. Human colorectal cancer cells frequently express IgG and display unique Ig repertoire. *World J Gastrointest Oncol*. 2019;11(3):195-207.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6425329/

12. https://www.uniprot.org/uniprot/Q01892

13. https://www.researchsquare.com/article/rs-136420/v1

Gene Function Analysis for Prostate Cancer Dataset (GSE71783)

Table 4: The Smallest Set of Genes Selected From Prostate Cancer Data by the Random Forest Method

Gene Symbol	Function
ACP5	Tartrate-Resistant Acid Phosphatase Type 5 It is involved in osteopontin/bone sialoprotein dephosphorylation and ferric iron binding. [1] Lyubimova et al. have observed that serum activity ACP5 in patients with breast cancer and prostate cancer having bone metastases is much higher than in healthy donors and patients without skeletal injuries. [2]
CENPBD1	CENPB DNA-Binding Domain Containing 1 Diseases associated with CENPBD1 include Mucopolysaccharidosis, Type Iva and Mucopolysaccharidosis Iv. [3]
MT1A	Metallothionein 1A Proteins encoded by this gene family serve as antioxidants and assist in detoxification of heavy metals. [4] Low expression of metallothionein proteins have been observed in many tumors. [5] Zinc plays an important role in prostate cancer pathogenesis and metallothionein are zinc-binding proteins. Previous studies suggest an association between metallothionein expression and prostate tumor progression [6]
PROM1	Prominin 1 It encodes a pentaspan transmembrane glycoprotein. It is often expressed on adult stem cells, where it is thought to maintain stem cell properties by suppressing differentiation. Expression of this gene is also associated with several types of cancer. [7] PROM1 was investigated as a prostate cancer stem cell marker using the same cell surface markers for identifying normal stem cells in the prostate. [8] Saha and colleagues have found that low levels of PROM1 expression are correlated with poor overall survival (OS) in prostate and lung cancers. [9]
QTRT1	Queuine TRNA-Ribosyltransferase Catalytic Subunit 1 This gene encodes the catalytic subunit of tRNA-guanine transglycosylase. [10] Sex Hormone Transfer RNAs play an important role in development and progression of breast and prostate cancer. [11] tRNA Q-modification has been associated with various forms of tumors. [12]

References:

1. https://www.uniprot.org/uniprot/P13686

2. Lyubimova, N.V., Pashkov, M.V., Tyulyandin, S.A. *et al.* Tartrate-resistant acid phosphatase as a marker of bone metastases in patients with breast cancer and prostate cancer. *Bull Exp Biol Med* 138, 77–79 (2004). https://link.springer.com/article/10.1007/BF02694481

3. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CENPBD1

4. https://www.ncbi.nlm.nih.gov/gene/4489

5. Cherian MG, Jayasurya A, Bay BH. Metallothioneins in human tumors and potential roles in carcinogenesis. Mutat Res. 2003 Dec 10;533(1-2):201-9. https://pubmed.ncbi.nlm.nih.gov/14643421/

6. Hlavna M, Raudenska M, Hudcova K, Gumulec J, Sztalmachova M, Tanhäuserova V, Babula P, Adam V, Eckschlager T, Kizek R, Kizek R, et al: microRNAs and zinc metabolism-related gene expression in prostate cancer cell lines treated with zinc(II) ions. Int J Oncol 41: 2237-2244, 2012 https://www.spandidos-publications.com/10.3892/ijo.2012.1655#b26-ijo-41-06-2237

7. https://www.ncbi.nlm.nih.gov/gene/8842

8. Collins AT, Berry PA, Hyde C, Stower MJ, Maitland NJ (2005) Prospective identification of tumorigenic prostate cancer stem cells. Cancer Res 65(23):10946–10951 https://pubmed.ncbi.nlm.nih.gov/16322242/

9. Saha, S.K., Islam, S.M.R., Kwak, KS. et al. PROM1 and PROM2 expression differentially modulates clinical prognosis of cancer: a multiomics analysis. Cancer Gene Ther 27, 147–167 (2020). https://www.nature.com/articles/s41417-019-0109-7

10. https://www.ncbi.nlm.nih.gov/gene/81890

11. Honda S, Loher P, Shigematsu M, et al. Sex hormone-dependent tRNA halves enhance cell proliferation in breast and prostate cancers. *Proc Natl Acad Sci U S A*. 2015;112(29):E3816-E3825. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517238/

12. Fergus C, Barnes D, Alqasem MA, Kelly VP. The queuine micronutrient: charting a course from microbe to man. *Nutrients*. 2015;7(4):2897-2929. Published 2015 Apr 15. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4425180/