EasyChair Preprint

№ 15681

# Gastrointestinal Cancer Classification by Symptomatology and Gene Expression Data Using Machine Learning

Tamanna Islam, Bithy Khanam, Fahmida Sultana, Kingkar Prosad Ghosh, Md. Kawsar Ahmed and Apurba Kumar Barman



January 7, 2025

2024 6th International Conference on Sustainable Technologies for Industry 5.0 (STI), 14-15 December, Dhaka

# Gastrointestinal Cancer Classification by Symptomatology and Gene Expression Data Using Machine Learning

Tamanna Islam[1], Bithy Khanam[1], Fahmida Sultana[1], Kingkar Prosad Ghosh[1,2], Md. Kawsar Ahmed[1], Apurba Kumar Barman[3]
[1]Department of Computer Science and Engineering, R. P. Shaha University, Narayanganj, Bangladesh
[2]Volgograd State Technical University, Volgograd, Russia
[3]Department of Pharmacy, R. P. Shaha University, Narayanganj, Bangladesh
Email: tamannaislam.20100026@rpsu.edu.bd, bithykhanam.20100053@rpsu.edu.bd, fahmidasultana.20100127@rpsu.edu.bd, kingkar@rpsu.edu.bd, kawsarahmedksp@gmail.com, apurba_phr@rpsu.edu.bd

***Abstract*— Cancer is the foremost reason for mortality around the globe, accounting for approximately ten million deaths annually. This multifaceted disease is differentiated by the unsupervised proliferation of unusual cells that can intrude adjacent tissues and propagate to remote body parts. This study concentrates on the classification of gastrointestinal (GI) cancers, specifically esophageal, liver, colon, stomach, and pancreatic cancers, due to their high incidence and mortality rates. Utilizing both symptomatology and gene expression data, we applied advanced machine learning (ML) techniques to enhance diagnostic accuracy for GI cancer classification. We developed an ML algorithm by integrating the best four classifiers (Support Vector Machine, Random Forest, Decision Tree, and Logistic Regression) and assessed its performance against ten distinct ML algorithms, such as Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Neural Network (NN), Gradient Boosting Machine (GBM), AdaBoost Classifier, and Extra Trees Classifier. Our proposed model exhibited top accuracy in classifying the five GI cancers in contrast to the other ML algorithms. The outcomes illustrate that ML algorithms can notably outperform traditional diagnostic methods, providing enhanced prognostic abilities. By utilizing symptom data and gene expression profiles, our approach emphasizes the evolving potential of ML in cancer diagnostics, paving the path for more precise and accurate medical interventions.**

***Keywords*— *Gastrointestinal Cancer, Cancer Classification, Machine Learning, Gene Expression, Liver Cancer, Esophageal Cancer, Stomach Cancer, Colon Cancer, Pancreatic Cancer.***

## I. INTRODUCTION

Globally, gastrointestinal (GI) carcinomas impact up to 4.8 million individuals annually, accounting for almost 25% of the worldwide cancer occurrence [1]. The most common types of GI cancers are: esophageal cancer, liver cancer, gastric (stomach) cancer, pancreatic cancer, colon cancer. Colon cancer, also known as colorectal cancer (CRC) and Gastric cancer (GC) are two kinds of malignant GI tumors with the largest incidence [2]. The GI organs are shown in Fig. 1.

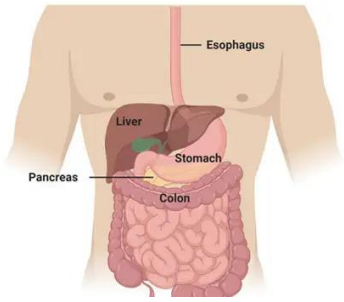


Fig. 1. GI Organs

Esophageal cancer (EC) predominantly develops in the lower esophagus and is associated with factors such as obesity, gastric reflux, and Barrett's esophagus [3]. Usually, it initiates in the inner wall cells of the esophagus and can occur anywhere along its length [4]. The liver processes food, digests nutrients, purifies blood, and detoxifies the body. Excess consumption of alcohol, exposure to toxins or drugs, diabetes, obesity, and viral hepatitis are all the risk factors for liver cancer [5]. Gastric Cancer (GC), also known as stomach cancer, is a malignant development of stomach cells that can begin anywhere in the stomach and spread to other organs, particularly the esophagus, lungs, and liver [6]. Pancreatic adenocarcinoma, primarily affecting the pancreas head, is caused by dangerous factors like smoking, family history, consuming alcohol, obesity, diabetes, and hypercholesterolemia [7]. Either the colon or the rectum is the source of the colorectal cancer (CRC); therefore, it can be identified as colon cancer or rectal cancer, depending on its point of origin [8]. According to the World Cancer Research Fund, colon cancer emerges as the 2nd leading cause of death from cancer. Liver cancer ranks is 3rd, stomach cancer stands 4th, esophageal cancer ranks 6th and pancreatic cancer concern as the 7th leading cause of cancer deaths globally.

This study aligns with Industry 5.0's vision by integrating artificial intelligence (AI) to enhance healthcare, promoting collaborative human-machine interactions in diagnostics, and enabling medical practitioners to make data-driven decisions. Our strategy focused on Industry 5.0 to improve medical diagnoses and empower medical practitioners to make better informed decisions that benefit patients. As a branch of AI, ML entails the development and application of algorithms to analyze data and its characteristics [9]. ML is increasingly

applied in cancer detection and research due to its advantages over traditional methods:

- Complex Pattern Recognition: ML can detect complex patterns in gene expression and symptoms that may not be visible in traditional diagnostics.
- Greater Accuracy and Efficiency: Learns from large datasets, boosting accuracy and speeding up diagnosis.
- Scalability and Adaptability: Easily retrained for various cancer types and new clinical data.
- Higher Precision and Speed: Provides real-time analysis without time-consuming lab tests.
- Reduced Human Error: Ensures consistent, reliable results.

The objective of this paper is to initially predict GI cancer by symptoms and then to classify the cancer using gene expression of the cancer cell. The following feature is present in this paper:

- We collected the GI cancer symptoms of raw data from the National Institute of Cancer Research & Hospital (NICRH) and created a dataset.
- We proposed a ML classification model to classify the liver, colon, pancreatic, esophageal, and stomach cancers using a novel approach.
- We trained the pre-trained or existing 10 models and our proposed model to compare the accuracy and precision of our proposed model.
- We identified a group of shared genes among malignancies originating from many GI locations that will assist in identifying routes and targets for therapy in GI malignancies.

## II. Related Work

In previous years, various ML methods were used in cancer detection and classification. Within this segment, we will go over a few relevant studies that improved our knowledge of cancer detection and classification methods.

In their work, they utilize ML to classify cancer genes from expression data. The proposed model enhances cancer gene classification using ML algorithms. The GBM are effective classifiers for cancer gene expression data in this paper. The accuracy obtained by other classifiers: SVM and GBM classifiers are 58.82% and 64.71% respectively [10].

In this paper, they identified cancer type based on gene expression data. They used SVM, K-Means, Formal Concept Analysis, and association rules for cancer classification. SVM achieved the highest accuracy of 99.8%. K-Means had an accuracy of 91.75%. The Formal Concept Analysis algorithm achieved an accuracy of 83.1%. SVM is the most accurate method for cancer classification [11].

The purpose of this paper is to classify esophageal and stomach cancers by using ML algorithms. They used DT, AdaBoost, and RF methods for classification. The DT model prone to overfitting, with inferior accuracy, AdaBoost is suitable for complex data types, improves performance through boosting, RF provides an effective method for classification imbalances, better than other algorithms [3].

This study focuses on applying ML techniques for colon cancer classification. They used the RF, DT, SVM, NB, and KNN models to classify colon cancer. Most of the models achieved over 95% accuracy for classification. The RF model had the highest accuracy of 96.8% and their proposed model achieved 96.8% accuracy [12].

In the study, the author illustrates two commonly used ML models, RF and Extreme Gradient Boosting (XGBoost), for classifying cancer data. The combination of XGBoost with the genetic algorithm achieved the highest accuracy score of 82% [13].

## III. Methodology

This section shows how the datasets were acquired, divided, preprocessed, and enhanced. Fig. 3 and Fig. 4 are the illustrations of our workflow with the models.

### A. Data Collection

We obtained GI cancer symptoms from the NICRH to build our own dataset. The NICRH is a prominent institution in Bangladesh dedicated to cancer care, research, and education. We collected 38 symptoms from 505 patients that were diagnosed with GI cancers (TABLE I). The dataset for GI cancer classification by using gene expression data (TABLE II) is collected from the Gene Expression Omnibus (GEO) site. The GEO site is a public repository and resource for gene expression data hosted by the National Center for Biotechnology Information (NCBI).

TABLE I. Symptom Dataset

| Attribute Description | Number | Dimension of dataset |
|---|---|---|
| Samples | 505 | 505 × 38 |
| Symptoms | 38 | |

TABLE II. Gene Expression Dataset

| Attribute Description | Number | Dimension of dataset |
|---|---|---|
| Samples | 802 | 802 × 5491 |
| Gene | 5491 | |

### B. Data Processing

The raw data of 5 different cancer symptoms from 505 patients were all integrated to produce a dataset for the initial diagnosis of GI cancer. The gene expression data that were collected from the GEO site was processed by removing all null values to create a single-valued dataset. In this study, the approach used for data normalization is min-max normalization, which is the method most frequently employed for data normalization. When preprocessing data to reduce the risk of overfitting, we also used Principal Component Analysis (PCA) for dimensionality reduction and cross-validation techniques to manage the complexity of gene expression data.

### C. Feature Selection

To go from a high-dimensional space to a latent space, we applied feature selection to both datasets. It provided a simple yet effective solution to this problem after removing redundant data. Additionally, we performed feature engineering in symptom dataset. The process of feature

engineering involves transforming unprocessed data into features in order to highlight pertinent information and improve ML models' capacity for data analytics [14].

### D. Data Split

We have divided the data into 70% for training, 20% for validation, and 10% for testing for both the symptom and gene expression dataset.

### E. Our Proposed Model

Our proposed model incorporates four ML algorithms—SVM, RF, DT, and LR—designed to better classify GI cancer. By combining these algorithms, our ensemble model mitigates the individual weaknesses and amplify strengths: SVM and RF provide high accuracy and robustness against overfitting, DT enhances interpretability, and LR offers probabilistic outputs. This layered approach outperforms single classifiers, creating a highly accurate, robust model for complex GI cancer datasets. Fig. 2 is the architecture of our proposed model.

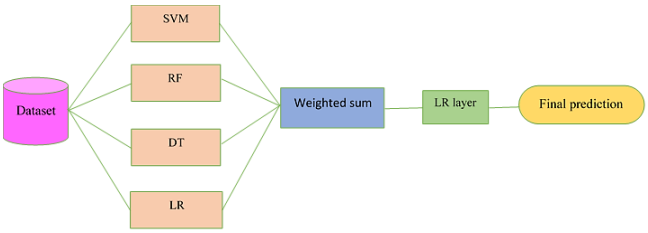


Fig. 2. Proposed Model Architecture

### F. Equation

From the base models, the outputs are used by our proposed model, denoted as $A_{SVM}$, $A_{RF}$, $A_{DT}$, and $A_{LR}$. These outputs act for the distinct probabilities or predictions from each classifier. The SVM classifies the data points, the RF promotes prediction accuracy, and it controls overfitting by using an ensemble of DT. The DT classifies the data based on feature splits. The LR predicts the probability of the target variable using a logistic function. Then the outputs are integrated and applied to the weighted sum. The logistic regression layer applies a weighted sum to these predictions to generate a final probability, given by:

$$Z = \alpha_0 + \alpha_1 A_{SVM} + \alpha_2 A_{RF} + \alpha_3 A_{DT} + \alpha_4 A_{LR} \quad (1)$$

$$A(y = 1 \mid X_{combined}) = 1 / (1 + e^{-Z}) \quad (2)$$

Here, Eq. (1) is the weighted sum. Eq. (2) is the cance of the target variable being 1 given the combined features. $A_{SVM}$, $A_{RF}$, $A_{DT}$, $A_{LR}$ are the predictions or probabilities predicted by each base model. $\alpha_0$, $\alpha_1$, $\alpha_2$, $\alpha_3$, $\alpha_4$ are the terms of the logistic regression model trained using combined features.

Our proposed model uses the supreme strengths of each base model. As a result, it gives more accurate and robust predictions in contrast to using any single model. The logistic regression layer acts as an integrator by combining these several predictors into one integrated and comprehensive final predictor. This method is particularly effective for handling complex and heterogeneous datasets. It considerably improves the overall classification performance for GI cancer detection.

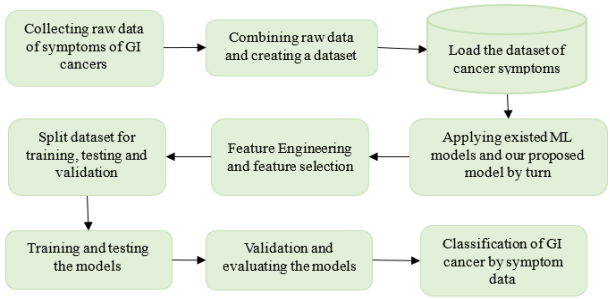


Fig. 3. Workflow of GI cancer classification by Symptom Data

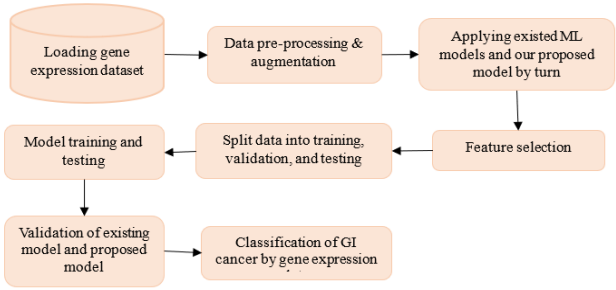


Fig. 4. Workflow of GI cancer classification by Gene Expression Data

## IV. Result and Analysis

In this section, we analyze gene characterization and related data to present experimental results and analysis of different ML models for GI cancer prediction.

### A. Performance Measurement for Cancer Classification

We analyze the performance of different ML models for classifying cancer types using both symptom and gene expression data. We analyze models based on four key metrics: accuracy, precision, recall, and F1 score, which offer a comprehensive view of each model's performance.

TABLE III and TABLE IV presents a comprehensive view of how different ML models perform in classifying cancer types based on symptomatology and gene expression data.

TABLE III. Symptom-Based Performance Measurement

| Symptom-Based | | | | |
|---|---|---|---|---|
| ***Model*** | ***Accuracy*** | ***Precision*** | ***Recall*** | ***F1score*** |
| SVM | 0.97 | 0.97 | 0.96 | 0.96 |
| LR | 0.91 | 0.90 | 0.89 | 0.89 |
| NB | 0.88 | 0.88 | 0.87 | 0.87 |
| DT | 0.86 | 0.89 | 0.86 | 0.86 |
| RF | 0.94 | 0.94 | 0.94 | 0.94 |
| KNN | 0.94 | 0.94 | 0.94 | 0.94 |
| NN | 0.96 | 0.96 | 0.96 | 0.96 |
| GBM | 0.96 | 0.97 | 0.96 | 0.96 |
| AdaBoost | 0.94 | 0.93 | 0.93 | 0.93 |
| Extra Trees Classifier | 0.97 | 0.96 | 0.96 | 0.96 |
| Our Proposed Model | 0.97 | 0.97 | 0.97 | 0.97 |

TABLE IV. GENE EXPRESSION-BASED PERFORMANCE MEASUREMENT

| Gene Expression-Based | | | | |
|---|---|---|---|---|
| ***Model*** | ***Accuracy*** | ***Precision*** | ***Recall*** | ***F1score*** |
| SVM | 0.96 | 0.98 | 0.98 | 0.98 |
| LR | 0.98 | 0.98 | 0.98 | 0.98 |
| NB | 0.97 | 0.96 | 0.96 | 0.96 |
| DT | 0.64 | 0.66 | 0.64 | 0.64 |
| RF | 0.97 | 0.98 | 0.98 | 0.98 |
| KNN | 0.94 | 0.96 | 0.96 | 0.96 |
| NN | 0.97 | 0.98 | 0.98 | 0.98 |
| GBM | 0.94 | 0.96 | 0.96 | 0.96 |
| AdaBoost | 0.94 | 0.96 | 0.96 | 0.96 |
| Extra Trees Classifier | 0.94 | 0.96 | 0.96 | 0.96 |
| Our Proposed Model | 0.99 | 0.99 | 0.99 | 0.99 |

For GI cancer classification, our proposed model illustrated distinguished performance in both gene expression and symptom datasets. In symptom-based GI cancer classification, our proposed model reached a high accuracy of 0.97, with adjusted precision, recall, and F1 score also at 0.97. This corresponds to the top performance of SVM, RF, and Extra Trees Classifier, which achieved accuracies up to 0.97 (Fig. 5), followed closely by NN and GBM at 0.96. This highlights the strong ability of the proposed model to adapt and perform better across different types of data.

Similarly for classification of GI cancers based on gene expression data, it achieved an exceptional accuracy of 0.99, with precision, recall, and F1 scores all at 0.99 (Fig. 6). This is better than individual models like LR, RF, and NN, which has accuracies between 0.94 to 0.98.

Our model achieved higher accuracy with gene expression data than with symptom-based classification. This is because gene expression profiles capture specific molecular signatures, making them more precise for differentiating GI cancer types, while symptom data is less specific and can overlap between cancers.

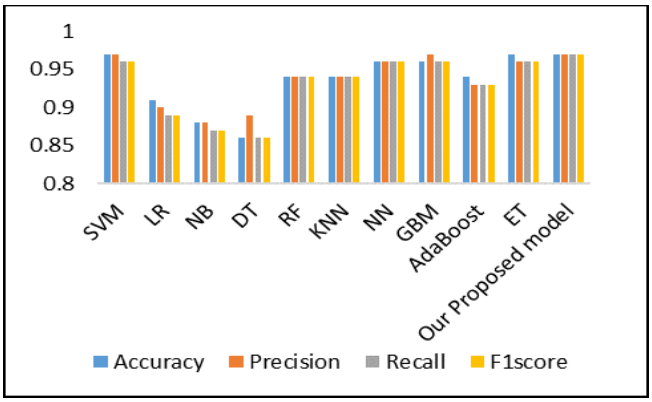


Fig. 5. Symptom-Based Performance Measurement

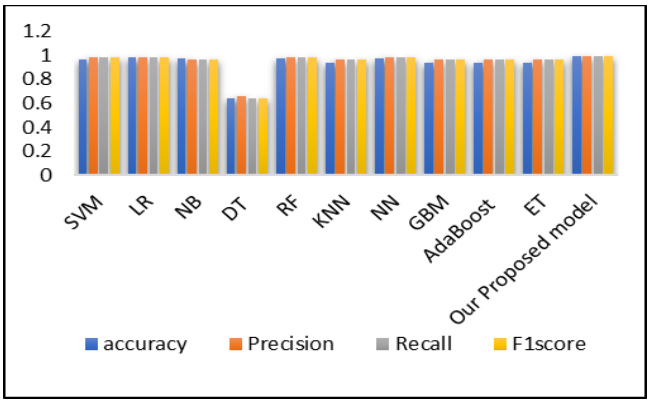


Fig. 6. Gene Expression-Based Performance Measurement

## B. *Comparison with Other Models*

TABLE V. PERFORMANCE COMPARISON WITH OTHER MODELS

| Model | Symptom-Based Accuracy | Gene Expression-Based Accuracy |
|---|---|---|
| SVM | 0.97 | 0.96 |
| LR | 0.91 | 0.98 |
| NB | 0.88 | 0.97 |
| DT | 0.86 | 0.64 |
| RF | 0.94 | 0.97 |
| KNN | 0.94 | 0.94 |
| NN | 0.96 | 0.97 |
| GBM | 0.96 | 0.94 |
| AdaBoost | 0.94 | 0.94 |
| Extra Trees Classifier | 0.97 | 0.94 |
| Our Proposed Model | 0.97 | 0.99 |

TABLE VI. TRAINING AND TESTING TIME COMPARISON WITH OTHER MODELS

| Models | Symptom-Based | | Gene Expression-Based | |
|---|---|---|---|---|
| | Training Time (sec) | Testing Time (sec) | Training Time (sec) | Testing Time (sec) |
| SVM | 0.5076 | 0.0582 | 3.1975 | 0.2202 |
| LR | 0.0514 | 0.0044 | 0.0993 | 0.0034 |
| NB | 0.0139 | 0.0052 | 0.0667 | 0.0157 |
| DT | 0.0187 | 0.0034 | 0.3292 | 0.0131 |
| RF | 0.6660 | 0.0187 | 0.0926 | 0.0030 |
| KNN | 0.0191 | 0.0129 | 0.0130 | 0.0449 |
| NN | 1.8817 | 0.0012 | 0.9209 | 0.0051 |
| GBM | 2.7037 | 0.0141 | 5.0919 | 0.0200 |
| AdaBoost | 1.4378 | 0.2985 | 2.7072 | 0.3469 |
| Extra Trees Classifier | 0.7686 | 0.0685 | 0.9945 | 0.0753 |
| Our Proposed Model | 0.0528 | 0.0009 | 0.3171 | 0.0039 |

We showed the performance of our proposed model and compare it with other models in TABLE V and TABLE VI represents only the training and testing time of the models. By combining gene expression and symptom data, our model outperformed top models (SVM, LR, DT, NB, RF, KNN, NN, and GBM), achieving 0.97 accuracy for symptom-based and 0.99 for gene expression-based classification.

### *C. Confusion Matrix*

An algorithm's performance, usually in supervised learning, can be shown using a particular table structure called a confusion matrix or error matrix [16]. This section discusses the confusion matrices for the symptom-based and gene expression-based recognition tests (Fig. 7 and Fig. 8). It is used to calculate the performance metrics of the models. The diagonal values represent True-Positive predictions.

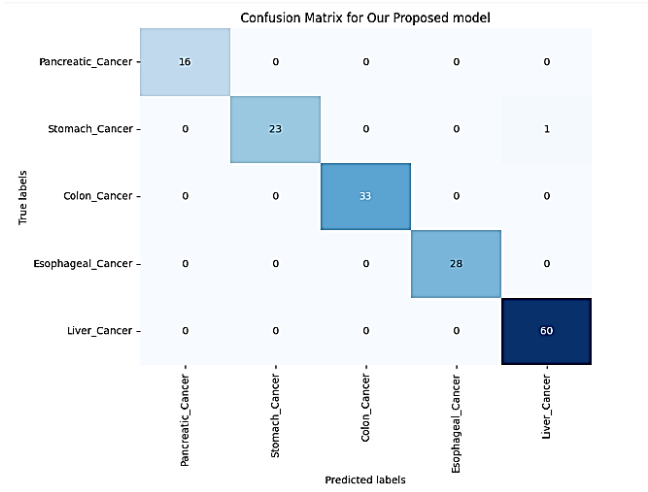


Fig. 7. Symptom-Based Confusion Matrix

In symptom-based classification (Fig. 7), the model accurately identified most cancers, with minor misclassifications involving liver and esophageal cancers. In gene expression-based classification (Fig. 8), it achieved near-perfect accuracy, with only one instance of stomach cancer misclassified as esophageal.

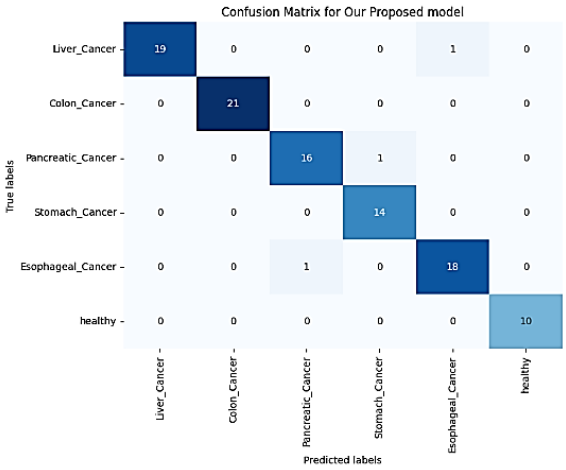


Fig. 8. Gene Expression-Based Confusion Matrix

### *D. Comparison with Previous Studies*

In this section, we present a comparative analysis of two recent studies on cancer classification, alongside our work (TABLE VII). This comparison highlights the approaches, models, and accuracies reported in the literature, illustrating where our model fits within the broader research landscape and demonstrating its effectiveness relative to existing methods.

TABLE VII. COMPARISON WITH PREVIOUS STUDIES

| Authors | Models and Methods | Accuracy (%) | Comments |
|---|---|---|---|
| Swain, Rashmi Ranjan, Debasish Swapnesh Kumar Nayak, and Tripti Swarnkar. [12] | RF, DT, SVM, Naive Bayes, and KNN | 96.8, 96.1, 95.2, 85.0, 95.0 respectively | This paper relies on a single dataset, which may not fully represent the variability of real-world colon cancer cases. |
| Kuntz, Sara, et al. [15] | CNN, whole-slide images (WSIs), Hazard Ratio (HR) | 84.5 to 90.4 respectively. | They primarily focus on gastrointestinal cancers, but their focus is mainly on colorectal and gastric cancers, with relatively limited attention to other types of GI cancers. |
| Our study | SVM, LR, NB, DT, RF, KNN, NN, GBM, AdaBoost, Extra Trees Classifier and our proposed model | Symptom-based: 0.97, 0.91, 0.88, 0.86, 0.94, 0.94, 0.96, 0.96, 0.94, 0.97, 0.97 respectively.<br><br>Gene expression-based: 0.96, 0.98, 0.97, 0.64, 0.97, 0.94, 0.97, 0.94, 0.94, 0.94, 0.99 respectively. | In our study, we have comprehensively discussed various GI cancers and analyzed the results using 10 different ML models and also our own proposed model. And the accuracy rate of our results is quite better than others. |

## V. Conclusion and Future Work

Our study demonstrates the crucial potential of ML techniques in the classification of gastrointestinal cancers, utilizing both symptom data and gene expression data. The ML model we proposed, which integrates four optimized classifiers (SVM, RF, DT, LR), has shown remarkable accuracy rates, achieving 97% accuracy in symptomatology-based classification and an impressive 99% accuracy in gene expression-based classification. These results underscore the superiority of ML algorithms over traditional diagnostic methods, highlighting their capacity to enhance diagnostic precision and prognostic accuracy.

The availability of relevant cancer data posed a challenge. Widening and diversifying the dataset could potentially increase accuracy rates. Since we had to rely on single institution data, the complexity of symptom and gene expression added further difficulties. However, our future study will focus on broadening the dataset and further enhancing the ML models to encompass more cancer types and a wider range of genetic profiles.

Our findings pave the way for the broader application of ML in cancer diagnostics, emphasizing its potential to enhance patient outcomes. By leveraging gene expression biomarkers for early detection, using symptom-based models for initial screening, and applying gene-based classification for personalized treatments, future cancer diagnostics can become more precise and effective. Additionally, the proposed model can be applied to various cancer types by retraining it with relevant symptom and gene expression data, allowing it to test and classify different cancers with high accuracy. This adaptability extends its diagnostic potential beyond GI cancers. The model's use of gene expression data enables precise cancer detection and subtype identification, supporting personalized treatment plans and informing decisions on targeted therapies. By advancing cancer diagnostics with sustainable technology, this work fosters a more efficient healthcare system, reducing resource use and promoting early interventions. ML-based diagnostics can streamline healthcare processes, building a resilient system that is better equipped to address global health challenges sustainably.

## References


[1] Arnold, M.; Abnet, C.C.; Neale, R.E.; Vignat, J.; Giovannucci, E.L.; McGlynn, K.A.; Bray, F. Global Burden of 5 Major Types of Gastrointestinal Cancer. Gastroenterology 2020, 159, 335–349.e15.

[2] Deng, S., Gu, J., Jiang, Z., Cao, Y., Mao, F., Xue, Y., Wang, J., Dai, K., Qin, L., Liu, K. and Wu, K., 2022. Application of nanotechnology in the early diagnosis and comprehensive treatment of gastrointestinal cancer. Journal of Nanobiotechnology, 20(1), p.415.

[3] Chen, Xiaoping, Lihui Zheng, Jianqi Yao, and Cheng-Fu Yang. "Classification of Esophageal Adenocarcinoma, Esophageal Squamous Cell Carcinoma, and Stomach Adenocarcinoma Based on Machine Learning Algorithms." Sensors & Materials 33 (2021).

[4] Hosseini, F., Asadi, F., Emami, H., & Ebnali, M. (2023). Machine learning applications for early detection of esophageal cancer: a systematic review. BMC Medical Informatics and Decision Making, 23(1), 124.

[5] Maurya, Bhawana, Saroj Hiranwal, and Manoj Kumar. "A review on liver cancer detection techniques." In 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), pp. 1-5. IEEE, 2020.

[6] Biglarian A, Hajizadeh E, Kazemnejad A, Zali M. "Application of artificial neural network in predicting the survival rate of gastric cancer patients." Iranian journal of public health 40.2 (2011): 80.

[7] Zhao, L., Zhao, H., & Yan, H. (2018). Gene expression profiling of 1200 pancreatic ductal adenocarcinoma reveals novel subtypes. BMC cancer, 18, 1-13.

[8] Sakr, A. S., Soliman, N. F., Al-Gaashani, M. S., Pławiak, P., Ateya, A. A., & Hammad, M. (2022). An efficient deep learning approach for colon cancer detection. Applied Sciences, 12(17), 8450.

[9] Tătaru, Octavian Sabin, Mihai Dorin Vartolomei, Jens J. Rassweiler, Oșan Virgil, Giuseppe Lucarelli, Francesco Porpiglia, Daniele Amparore et al. "Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives." Diagnostics 11, no. 2 (2021): 354.

[10] Joseph, M., Devaraj, M., & Vea, L. A. (2018, November). Cancer classification of gene expression data using machine learning models. In 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM) (pp. 1-6). IEEE.

[11] He, Yinchao, Ryan Bockmon, Miracle Modey, and Sarah Roscoe. "Classification of cancer types based on gene expression data." In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2175-2182. IEEE, 2020.

[12] Swain, Rashmi Ranjan, Debasish Swapnesh Kumar Nayak, and Tripti Swarnkar. "A Comparative Analysis of Machine Learning Models for Colon Cancer Classification." In 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT), pp. 1-5. IEEE, 2023.

[13] Desdhanty, Vabiyana Safira, and Zuherman Rustam. "Liver cancer classification using random forest and extreme gradient boosting (XGBoost) with genetic algorithm as feature selection." In 2021 International conference on decision aid sciences and application (DASA), pp. 716-719. IEEE, 2021.

[14] Alharbi, Fadi, and Aleksandar Vakanski. "Machine learning methods for cancer classification using gene expression data: A review." Bioengineering 10.2 (2023): 173.

[15] Kuntz, S., Krieghoff-Henning, E., Kather, J.N., Jutzi, T., Höhn, J., Kiehl, L., Hekler, A., Alwers, E., von Kalle, C., Fröhling, S. and Utikal, J.S., 2021. Gastrointestinal cancer classification and prognostication from histology using deep learning: Systematic review. European Journal of Cancer, 155, pp.200-215.

[16] Nodi NI, Ghosh KP, Ferdousi Z. Machine Learning Approaches for Classifying Online Job Advertisements. RPSU RESEARCH JOURNAL.:38.