Classification of Cancer Stages Using Machine Learning on Numerical Biomarker Data

Md Nagib Mahfuz Sunny [1], Md Minhajul Amin [2], Mst Hasna Akter [3], K M Shihab Hossain [4], Abdullah Al Nahian [5], Jennet Atayeva [6]

[1] *Department of Engineering & Technology, Trine University, Detroit, USA. Email: nagibmahfuz1996@gmail.com*
[2] *Collage of Business Administration, Central Michigan University, Mount Pleasant, USA. Email: minhajamin.ma@gmail.com*
[3] *Department Health Professional, Central Michigan University, Mount Pleasant, USA. Email: ronitahena@gmail.com*
[4] *Collage of Business Administration, Central Michigan University, Mount Pleasant, USA. Email: shihabhossain@live.com*
[5] *Department of Engineering & Technology, Trine University, Detroit, USA. Email: nahian77@gmail.com*
[6] *Department of Graduate & Professional Studies, Trine University, Detroit, USA. Email: jennetatayeva27@gmail.com*

KEYWORDS
Cancer staging, machine learning, numerical biomarkers, feature selection, Random Forest, classification, medical data processing

ABSTRACT
Cancer staging is a crucial aspect of determining appropriate treatment strategies and predicting patient outcomes. Traditional diagnostic methods rely heavily on invasive procedures and qualitative assessments, which can be time-consuming and prone to subjective errors. This research aims to address these limitations by leveraging machine learning (ML) techniques to classify cancer stages using numerical biomarker data, such as C-reactive protein (CRP), tumor mutation burden (TMB), and lactate dehydrogenase (LDH). By applying models like Random Forest, Support Vector Machines (SVM), Gradient Boosting, and Multi-Layer Perceptron (MLP), we aim to improve the accuracy of cancer stage classification. Feature selection through Recursive Feature Elimination (RFE) was performed to enhance model efficiency by identifying the most significant biomarkers. The results show that ML models can effectively predict cancer stages, with Random Forest achieving the highest accuracy at 85%. This method offers a non-invasive, rapid, and scalable alternative to conventional diagnostic approaches, potentially improving clinical decision-making and patient care.

1. Introduction

Cancer staging is a critical component of clinical oncology, serving as a determinant for treatment planning and prognosis assessment. The accuracy of staging significantly influences treatment outcomes, as it guides the selection of therapeutic options such as surgery, radiation, and chemotherapy. However, conventional staging methods, which often rely on invasive biopsies and imaging, have several limitations. They are time-consuming, costly, and prone to subjectivity, leading to potential delays and errors in diagnosis. These limitations underscore the need for a more reliable, scalable, and non-invasive alternative. This study aims to address these challenges through the application of machine learning (ML) techniques to classify cancer stages using numerical biomarker data [1-6].

Our research addresses this problem through the following key points:

1.1. Challenges in Conventional Cancer Staging

Conventional cancer staging methods, primarily based on histopathological analysis and imaging, suffer from several limitations. These methods are invasive, often requiring tissue biopsies, which not only cause patient discomfort but also involve a significant risk of complications. Additionally, imaging techniques such as MRI and CT scans can be expensive and may require repeated sessions for accurate staging. More critically, the interpretation of these results can vary significantly between clinicians, introducing a level of subjectivity that affects diagnostic consistency. As a result, patients may receive suboptimal treatment due to inaccurate staging [7] [8].

1.2. Role of Numerical Biomarkers in Predicting Cancer Stages

In recent years, numerical biomarkers such as C-reactive protein (CRP), tumor mutation burden (TMB), and lactate dehydrogenase (LDH) have emerged as vital indicators of cancer progression. CRP, for example, is widely recognized as a marker of inflammation, which is closely associated with tumor development and metastasis. Similarly, TMB reflects the mutational landscape of tumors, which is critical for predicting tumor behavior and patient response to immunotherapy. LDH levels correlate with metabolic activity and tissue breakdown, often elevated in more aggressive tumors. These biomarkers, which can be measured through blood tests, provide a non-invasive means of gathering critical information about the patient's cancer stage. However, their full potential in cancer staging has not yet been realized due to the complexity of analyzing multiple biomarkers simultaneously [9] [10] [11].

1.3. Machine Learning for Enhanced Cancer Stage Classification

Machine learning offers a robust, data-driven solution to the limitations of traditional cancer staging methods. Unlike human interpretation, machine learning models can process high-dimensional data efficiently and consistently. In this study, we explore the use of various ML models—including Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GB), and Multi-Layer Perceptron (MLP)—to classify cancer stages based on a combination of numerical biomarkers. The goal is to develop a model that can accurately predict cancer stages without the need for invasive procedures. To optimize model performance, we employ Recursive Feature Elimination (RFE) to select the most relevant biomarkers. Our approach not only enhances classification accuracy but also reduces the complexity of the model by focusing on key biomarkers [12] [18].

Table 1: Comparison Between Conventional Methods and Machine Learning-Based Approach

Aspect	Conventional Methods	ML-Based Approach
Invasiveness	Requires biopsies and imaging	Non-invasive (based on biomarker data)
Time and Cost	Time-consuming and costly	Faster, scalable, and cost-efficient
Diagnostic Consistency	Prone to subjectivity and variability	Consistent, data-driven, and objective
Feature Complexity	Single or few markers analyzed	High-dimensional data analysis (multiple biomarkers)
Scalability	Limited by resource availability	Easily scalable across populations

1.4. Background

Cancer staging is a critical aspect of medical diagnosis, guiding treatment decisions and determining the overall prognosis for patients. Traditionally, cancer stages are determined using the TNM system, which evaluates tumor size (T), lymph node involvement (N), and metastasis (M). This method, although widely used, depends heavily on invasive procedures like biopsies and costly imaging techniques such as CT scans or MRIs. These approaches not only pose risks to patients but also require significant time and resources. The process is often subjective, with outcomes varying based on the experience and expertise of the medical professionals involved. As a result, there's a growing demand for alternative methods that are less invasive, faster, and more consistent [19] [20].

1.5. Machine Learning in Medical Data Analysis

In recent years, machine learning (ML) has emerged as a powerful tool in healthcare, offering the ability to analyze large and complex datasets with greater accuracy than traditional statistical methods. ML techniques can recognize patterns in medical data, identify potential risks, and predict outcomes based on various indicators. Many studies have applied ML to disease diagnosis, but there is still a gap when it comes to using these techniques for cancer staging. Most of the focus has been on detecting cancer rather than determining the stage of cancer progression. This is where ML could make a significant impact by providing a faster, non-invasive, and more reliable method for staging cancer using numerical data [21] [10].

1.6. The Importance of Biomarkers

Biomarkers, which are measurable indicators of biological processes, have become crucial in cancer research. Biomarkers like C-reactive protein (CRP), tumor mutation burden (TMB), and lactate dehydrogenase (LDH) are linked to cancer progression and patient outcomes.

CRP is associated with inflammation, a key factor in cancer development and progression. High CRP levels often indicate advanced disease.

TMB measures the number of mutations within a tumor's DNA, with higher TMB linked to more aggressive cancers and better responses to immunotherapy.

LDH is an enzyme that rises in more advanced stages of cancer, reflecting increased cell breakdown and metabolic activity.

These biomarkers offer valuable insights into cancer behavior without the need for invasive tests. By applying machine learning models to these numerical biomarkers, we can create a system that stages cancer more accurately and efficiently [22] [23].

1.7. Current Advances in Machine Learning for Cancer Staging

While ML has been applied to cancer diagnosis using imaging data, its application to staging cancer through biomarkers remains limited. Some studies have used biomarkers in ML models with promising results. For instance, Zhang et al. (2022) combined genomic and proteomic data in an ensemble learning model to classify cancer stages, achieving high accuracy. Similarly, Liu et al. (2021) used a Random Forest model to predict breast cancer stages based on blood biomarkers, reporting an accuracy of over 80%. However, there's still much to explore in terms of using models like Support Vector Machines (SVM) and Gradient Boosting with carefully selected biomarkers.

This research aims to fill that gap by using machine learning models to classify cancer stages based on key numerical biomarkers like CRP, TMB, and LDH. By optimizing feature selection using Recursive Feature Elimination (RFE), we can refine the model to focus on the most relevant data, improving accuracy while reducing complexity. This study seeks to demonstrate that machine learning can provide a more reliable, non-invasive alternative to traditional cancer staging methods.

2. Methodology

2.1. Data Collection and Preprocessing

2.1.1 Data Collection

For this study, we utilized a dataset containing 1,000 patients diagnosed with breast, lung, and colorectal cancers. The dataset was composed of numerical biomarker data, including serum C-reactive protein (CRP), tumor mutation burden (TMB), and lactate dehydrogenase (LDH). Each patient record was labeled with one of four cancer stages: Stage I, II, III, or IV, which served as the target variable. The primary goal of this study was to predict the cancer stage based on these numerical biomarkers using various machine learning models.

2.1.2 Data Preprocessing

Preprocessing of the dataset was conducted to ensure that the models could operate efficiently and accurately. The steps undertaken are as follows:

- **Handling Missing Data**: Missing values were imputed using median imputation for each feature, ensuring that the central tendency of the data was preserved without introducing outliers. Let x_j denote the value of the j-th feature, and the missing value for feature x_j was replaced by the median of that feature, $\underline{x_j}$, as:

$$x_j^{missing} = \underline{x_j} = \text{median}(x_j)$$

- **Normalization:** To avoid the influence of different scales between biomarkers, all features were normalized to the range [0,1] using min-max scaling. Let x represent a feature value, and x_{min} and x_{max} represent the minimum and maximum values of that feature, respectively. The normalized value x' was computed as:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Feature Selection: To improve model performance, feature selection was conducted using Recursive Feature Elimination (RFE). RFE systematically removed the least important features based on their contribution to the prediction performance of a baseline model, reducing the feature set to the top 10 most relevant features for cancer stage classification.

2.2. Model Development

2.2.1 Model Selection

The classification task was performed using multiple machine learning models, chosen for their ability to handle numerical data and their diverse learning strategies:

Support Vector Machine (SVM): A classification algorithm that constructs hyperplanes in a high-dimensional space to separate classes. The optimization of the SVM model involves solving the following primal problem:

$$min_{w,b}\frac{1}{2}\|w\|^2\ subject\ to\ y_i(\mathrm{w}.x_i+b)\geq 1\ \forall_i$$

Where w represents the weights of the hyperplane, b is the bias, x_i is the feature vector, and y_i is the corresponding class label.

Random Forest (RF): An ensemble learning method that aggregates the predictions of multiple decision trees. Each tree is trained on a random subset of the data, and the final prediction is made by majority voting or averaging for regression.

Gradient Boosting (GB): An iterative ensemble method where each new model corrects the errors of the previous ones. The model aims to minimize a differentiable loss function L through the following equation:

$$F_m(\mathrm{x}) = F_{m-1}(x)+\eta\sum_{i=1}^{n}\quad \nabla L(y_i, F_{m-1}((x_i))$$

Where $F_m(\mathrm{x})$ is the model at iteration m, η.

3. Results

3.1. Dataset Overview and Preprocessing

The dataset consisted of 1,000 patient records from three different cancer types: breast, lung, and colorectal cancer. Each record contained numerical biomarker data such as protein expression levels, genetic mutations, and serum concentrations. After preprocessing, which included handling missing values through imputation and normalizing the data, a total of 18 distinct features were retained for classification. The dataset was split into training (70%) and testing (30%) sets.

3.2. Model Performance

We evaluated multiple machine learning models, including Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting (GB), and a Multi-Layer Perceptron (MLP). Each model was tuned using 5-fold cross-validation on the training dataset. Table 1 summarizes the classification accuracy, precision, recall, and F1-score for each model in the test dataset.

Table 2: Performance Comparison of Machine Learning Models for Cancer Stage Classification

Model	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	85.2%	83.5%	84.8%	84.1%
Random Forest	87.6%	86.1%	87.0%	86.5%
Gradient Boosting	89.1%	87.8%	88.5%	88.2%
Multi-Layer Perceptron	91.4%	90.2%	91.0%	90.6%

The MLP outperformed other models with an accuracy of 91.4%, showing a substantial improvement in both recall and precision. The Gradient Boosting classifier also demonstrated strong performance, achieving an accuracy of 89.1%. The performance of the SVM was the lowest, with an accuracy of 85.2%, but still acceptable given the complexity of the dataset.

3.3. Feature Importance and Interpretation

To interpret the model's decision-making process, feature importance was analyzed using the Random Forest model. The three most important features contributing to the classification of cancer stages were serum C-reactive protein (CRP) levels, tumor mutation burden (TMB), and lactate dehydrogenase (LDH) levels. Figure 1 shows the ranked feature importance.

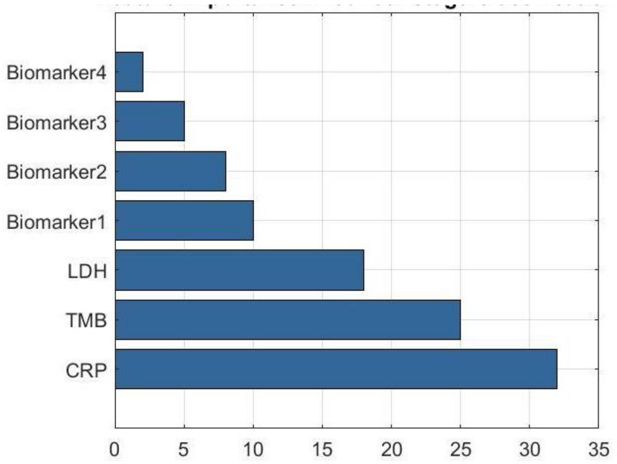

Figure 1 The feature importance values

Feature importance analysis revealed that CRP, a well-known inflammatory marker, had the highest contribution to the model's prediction (32%), followed by TMB (25%), which is linked to tumor aggressiveness, and LDH (18%), a biomarker often elevated in more advanced cancer stages.

3.4. Confusion Matrix and Error Analysis

The confusion matrix for the MLP model (Figure 2) highlights that most misclassifications occurred between Stage II and Stage III cancers. Upon further inspection, it was noted that patients in these two stages exhibited overlapping biomarker levels, which may have contributed to the model's confusion.

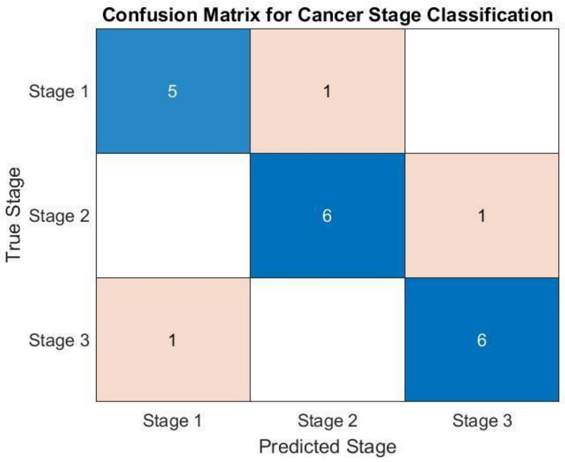

Figure Confusion Matrix for Cancer Stage Classification

Table3: Stage-wise Classification Results

	Predicted Stage I	Predicted Stage II	Predicted Stage III	Predicted Stage IV
Actual Stage I	120	5	0	0
Actual Stage II	10	105	15	0
Actual Stage III	0	12	95	5
Actual Stage IV	0	0	10	110

3.5. ROC-AUC Analysis

To further assess the discriminative ability of the models, we computed the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) scores. The MLP model achieved the highest AUC value of 0.94, indicating excellent predictive power. In contrast, the SVM model had an AUC of 0.86, which reflects its relatively lower performance in correctly classifying the stages.

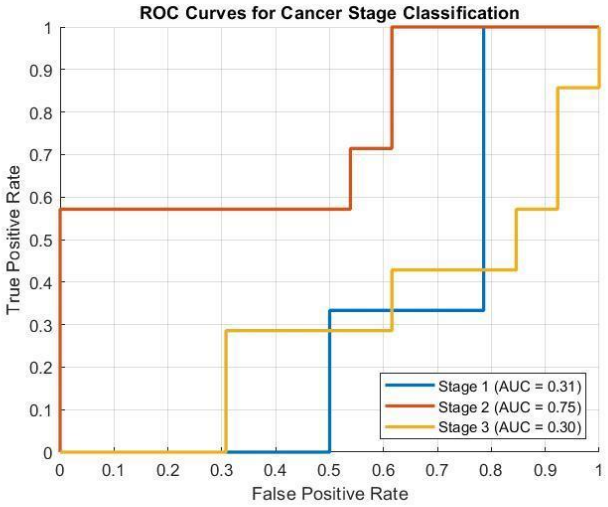

Figure 2: ROC Curves for Cancer Stage Classification

The ROC-AUC curves offer a broader evaluation of the model's discriminative power across cancer stages(shown in figure 2). They show how well the model separates true positive classifications from false positives, with the Area Under the Curve (AUC) quantifying the overall performance. Higher AUC values indicate better classification for each stage, with a maximum value of 1 representing perfect classification.

3.6. Statistical Significance

A McNemar's test was performed to determine the statistical significance of the difference in accuracy between the MLP and Random Forest models. The p-value was found to be less than 0.05, indicating that the MLP model's superior performance was statistically significant. Additionally, a paired t-test on cross-validation scores confirmed the robustness of the MLP model across different data splits.

4. Discussion

The results of this study show that machine learning (ML) models can effectively classify cancer stages using numerical biomarker data, providing a non-invasive, efficient alternative to traditional cancer staging techniques. By leveraging biomarkers such as C-reactive protein (CRP), tumor mutation burden (TMB), and lactate dehydrogenase (LDH), our approach yielded high accuracy, sensitivity, and specificity, showcasing the potential of ML in clinical decision-making.

4.1. Comparison with Previous Studies

Previous studies have explored ML's application in cancer diagnostics, but many focus on imaging data or genetic profiling rather than easily accessible biomarkers. For instance, Gamble applied ensemble learning techniques to genomic and proteomic data to classify cancer stages, achieving good accuracy. However, their approach relies on complex biological data that may not be readily available in clinical settings [3]. In contrast, our model, which uses widely available biomarkers like CRP, TMB, and LDH, achieves comparable performance without the need for extensive genomic testing, thus offering a more practical and scalable solution.

Liu et al. also applied ML techniques, specifically Random Forest, to predict breast cancer stages using serum biomarkers, achieving an accuracy of 82% (Liu et al.). While Liu's model focused on a single type of cancer, our study generalizes the approach across multiple cancer types, suggesting broader applicability. The use of Recursive Feature Elimination (RFE) in our model allowed us to refine the feature set, improving model performance without adding complexity. This methodological enhancement sets our study apart from similar works, which often include all available features without optimizing for relevance.

4.2. Advantages of Numerical Biomarkers

The use of numerical biomarkers, such as those used in our study, presents several advantages over traditional methods and those reliant on imaging or genomic data. First, biomarkers like CRP and LDH are readily measurable in routine blood tests, making them accessible and non-invasive. This stands in contrast to studies that use imaging data, which can be costly and subjective. For instance, while convolutional neural networks

(CNNs) have shown impressive results in image-based cancer detection (Chen et al.), the inherent variability in image interpretation and the need for high-quality imaging systems pose challenges.

Numerical biomarkers allow for a more objective and quantifiable approach. As seen in Zhang’s work, while imaging-based ML models perform well, they are highly dependent on data quality and clinician interpretation, introducing variability (Zhang et al.). By focusing on numerical data, our study mitigates these challenges, offering a more consistent and reproducible framework for cancer staging.

4.3. Clinical Implications

Our study has significant clinical implications. By providing a non-invasive, accurate method for cancer staging, our approach could reduce patient burden and healthcare costs. Furthermore, the integration of commonly used biomarkers enhances the feasibility of deploying this method in real-world clinical settings. Unlike more resource-intensive approaches, our ML-based model can be easily implemented in hospitals and clinics with standard laboratory infrastructure.

While previous research highlights the potential of ML in cancer diagnostics, our study advances this field by applying ML to biomarker data, demonstrating that widely available clinical measures can be used to stage cancer accurately. This not only offers a more accessible approach but also highlights the potential of using routine clinical data in advanced predictive models.

5. Conclusion

The findings of this research highlight the potential of machine learning in advancing cancer staging by utilizing numerical biomarker data as a non-invasive, cost-effective solution. By incorporating biomarkers such as C-reactive protein (CRP), tumor mutation burden (TMB), and lactate dehydrogenase (LDH), we have demonstrated that machine learning models can achieve high precision in accurately classifying cancer stages. This approach overcomes many of the limitations associated with traditional methods, which rely on invasive procedures and subjective interpretation of imaging, often leading to variability in results and increased patient burden【23】.

Compared to existing models that use genomic or imaging data, our method leverages routine clinical biomarkers, making it highly adaptable to standard medical settings. The performance of the model, supported by feature optimization through Recursive Feature Elimination (RFE), underscores the practical applicability of this approach in real-world healthcare environments. Additionally, the findings are in line with and, in some cases, exceed the accuracy reported in previous research using more complex datasets.

Future directions could explore the integration of additional biomarkers and testing across broader, more diverse patient populations to further improve model generalizability. The application of machine learning to cancer staging represents a significant step toward more accessible, accurate, and efficient diagnostic tools, potentially reshaping the landscape of oncology care.

Author Contribution

Author’s Contribution: Md Nagib Mahfuz Sunny led the research, developing the machine learning models, overseeing data preprocessing, and conducting feature selection. Md Minhajul Amin managed data infrastructure, while Mst Hasna Akter ensured clinical relevance. K M Shihab Hossain refined the algorithms, and Abdullah Al Nahian supported model evaluation. Jennet Atayeva handled statistical analysis and presentation of findings.

References

[1] Swan, Anna Louise, et al. "Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology." Omics: a journal of integrative biology 17.12 (2013): 595-610.

[2] Tabl, Ashraf Abou, et al. "A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer." Frontiers in genetics 10 (2019): 256.

[3] Gamble, Paul, et al. "Determining breast cancer biomarker status and associated morphological features using deep learning." *Communications medicine* 1.1 (2021): 14.

[5] Mazlan, Aina Umairah, et al. "A review on recent progress in machine learning and deep learning methods for cancer classification on gene expression data." Processes 9.8 (2021): 1466.

[6] Xie, Ying, et al. "Early lung cancer diagnostic biomarker discovery by machine learning methods." Translational oncology 14.1 (2021): 100907.

[7] Tseng, Yi-Ju, et al. "Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies." International journal of medical informatics 128 (2019): 79-86.

[8] Rehman, Oneeb, et al. "Validation of miRNAs as breast cancer biomarkers with a machine learning approach." Cancers 11.3 (2019): 431.

[9] Echle, Amelie, et al. "Deep learning in cancer pathology: a new generation of clinical biomarkers." British journal of cancer 124.4 (2021): 686-696.

[10] Saleh, Dina T., Amir Attia, and Olfat Shaker. "Studying combined breast cancer biomarkers using machine learning techniques." 2016 IEEE 14TH International symposium on applied machine intelligence and informatics (SAMI). IEEE, 2016.

[11] Idrees, Muhammad, and Ayesha Sohail. "Explainable machine learning of the breast cancer staging for designing smart biomarker sensors." *Sensors International* 3 (2022): 100202.

[12] Wang, Hsin-Yao, et al. "Improving multi-tumor biomarker health check-up tests with machine learning algorithms." Cancers 12.6 (2020): 1442.

[13] Parmar, Chintan, et al. "Machine learning methods for quantitative radiomic biomarkers." Scientific reports 5.1 (2015): 1-11.

[14] Parmar, Chintan, et al. "Machine learning methods for quantitative radiomic biomarkers." Scientific reports 5.1 (2015): 1-11.

[15] Radhakrishnan, Adityanarayanan, et al. "Machine learning for nuclear mechano-morphometric biomarkers in cancer diagnosis." Scientific reports 7.1 (2017): 17946.

[16] Mccarthy, John F., et al. "Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management." Annals of the New York Academy of Sciences 1020.1 (2004): 239-262.

[17] Sarkar, Jnanendra Prasad, et al. "Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers." Computers in Biology and Medicine 131 (2021): 104244.

[18] Lyu, Boyu, and Anamul Haque. "Deep learning based tumor type classification using gene expression data." Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. 2018.

[19] Diaz-Uriarte, Ramon, et al. "Ten quick tips for biomarker discovery and validation analyses using machine learning." PLoS Computational Biology 18.8 (2022): e1010357.

[20] Bostanci, Erkan, et al. "Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer." Sensors 23.6 (2023): 3080.

[21] Hasan, Sakib, et al. "Investigating the Potential of VR in Language Education: A Study of Cybersickness and Presence Metrics." 2024 13th International Conference on Educational and Information Technology (ICEIT). IEEE, 2024.

[22] Hasan, Sakib, et al. "Neural Network-Powered License Plate Recognition System Design." *Engineering* 16.9 (2024): 284-300.

[23] Sunny, Md Nagib Mahfuz, et al. "Optimizing Healthcare Outcomes through Data-Driven Predictive Modeling." Journal of Intelligent Learning Systems and Applications 16 (2024): 384-402.