dATTA_P2

by manoj v

Submission date: 30-Jul-2025 05:52PM (UTC+0530)
Submission ID: 2722798454
File name: Datta_report_2_final_2_2.docx (246.21K)
Word count: 5502
Character count: 33585

Detection Of Stomach Adenocarcinoma Using Ensemble Machine Learning Framework

Given Name Surname
dept. name of organization
(of Affiliation)
name of organization
City, Country
email address or ORCID

***Abstract*—Stomach adenocarcinoma (STAD) is one of the most aggressive gastrointestinal cancers, often diagnosed at advanced stages due to the lack of early diagnostic markers. This study proposes an integrative machine learning (ML) framework for STAD classification using a combination of miRNA expression, RNA-Seq profiles, and clinical metadata. Data was obtained from The UCSC Xena and carefully curated, pre-processed, and merged, including encoding categorical clinical features. Various ensemble machine learning models, which includes Bagging, Random Forest, Voting, Stacking and etc were trained and evaluated with key metrics. Among all these model, Voting Classifier (RF+LR+CatB) have achieved the highest performance with a accuracy of 98.77%, a AUPR of 0.9922 and a ROC-AUC scores of 0.9955. The findings suggest that the approach can help researchers and doctors by providing more reliable prediction tools to support diagnosis and treatment planning.**

Keywords— STAD, miRNA, mRNA, Machine learning, ensemble learning, Gene Expression, Accuracy, clinical, Gene Expression.

I. INTRODUCTION

Gastric cancer remains one of the leading causes of cancer-related mortality worldwide, with low survival rates due to late diagnosis and limited treatment efficacy [1]. Precise prognosis prediction is important for individualized treatment planning and enhancing patient outcomes. MicroRNAs (miRNAs) are key post-transcriptional regulators of cancer development and progression, holding promise as valid prognostic biomarkers [2], [3].

Advances in molecular biology and high-throughput tools like RNA sequencing and microarray analysis have made it possible to obtain gene expression profiles from tumor samples. This data provides important insights into the biological processes involved and can lead to better diagnosis, classification, and personalized treatment plans for stomach cancer.

Current advances in high-throughput sequencing and machine learning (ML) have enabled one to integrate genomic and clinical information for cancer survival prediction. In particular, ensemble machine learning models, which combine multiple base learners to improve predictive performance, have been demonstrated to outperform a single classifier in various biomedical applications [4], [5]. Among the ensemble methods, stacking, voting, blending, and bagging have been widely used due to their robustness, ability to detect non-linear relationships, and ability to avoid overfitting [6].

We evaluated miRNA expression profiles, mRNA-Seq gene expression values and clinical information of the TCGA-STAD (Stomach Adenocarcinoma) dataset for finding prognostic biomarkers and constructing ensemble ML models to predict overall survival (OS). A comprehensive set of base classifiers including Random Forest (RF), Logistic Regression (LR), XG Boost, Light GBM, Cat Boost, MLP, Ada Boost, and Quadratic Discriminant Analysis were utilized. Ensemble techniques such as soft voting, stacking (with various meta-learners), balanced random forests, and hybrid meta-ensembles were explored and compared based on metrics including accuracy, ROC AUC, F1 score, and AUPR.

Our results demonstrated that ensemble strategies, mainly soft voting and stacking, significantly outperformed individual classifiers. Notably, the Voting Classifier combining RF, LR, and Cat Boost achieved the highest Accuracy and ROC AUC values, highlighting the potential of ensemble learning in cancer prognosis modeling.

Combining this high-dimensional genomic data with machine learning enables predictive models capable of differentiating between cancer and non-cancer samples with remarkable accuracy. These models also shed light on the most informative genes that contribute to disease classification. This reduces not just the ability to enhance early diagnosis but also sets the stage for individualized treatment approaches.

II. RELATED WORK

The last few years have witnessed a remarkable increase in the utilization of machine learning and bioinformatics strategies for Stomach cancer survival prediction and the detection of biomarkers, especially from miRNA and transcriptomics data. Ruan et al. [7] and Haider et al. [8] highlighted the utility of miRNAs as diagnostic and prognostic markers because of their stability, regulatory role, and differential expression among cancers. The TCGA-STAD initiative has facilitated large-scale investigation of these patterns through well-annotated molecular and clinical datasets.

Stacking, first proposed by Wolpert [11], has benefited genomic research by aggregating predictions of heterogeneous classifiers into an upper-level meta-learner. It improves performance by minimizing bias and variance and has been applied to multi-omics integration problems. For dealing with class imbalance in survival data, methods like the Balanced Random Forest [6] and sampling methods like SMOTE have been utilized in cancer genomics.

Studies by Liu et al. [14] and Zhang et al. [15] demonstrated that ensemble learning combined with miRNA expression data could accurately predict clinical outcomes in gastrointestinal cancers, including STAD. Integrative

frameworks that combine miRNA, mRNA, and clinical metadata have been shown to outperform single-modal models in terms of generalization and clinical relevance.

While past studies have shown encouraging results, several important gaps still remain. Many models tend to focus on a limited set of features or explore only one type of ensemble method, which makes it difficult to compare their effectiveness or apply them more broadly. Additionally, only a few efforts have tried to benchmark a wide range of ensemble techniques using consistent miRNA and clinical data. This study aims to fill those gaps by evaluating different ensemble machine learning models—including Voting, Stacking, Bagging, and Boosting—on a carefully prepared dataset. comparison of model performance and uncover meaningful insights into potential biomarkers for predicting patient survival.

TABLE I. Summary of Related Work

Author	**Method**	**Key Features**	**Performance / Results**	**Limitations**
Ruan et al. [7]	Graph Neural Networks (GNNs)	miRNA-disease associations	High AUC, effective in capturing network-based relationships	Limited to association prediction; not survival-focused
Haider et al. [8]	Comparative ML models (SVM, RF, XGBoost)	Gene expression, miRNA	RF & XGBoost performed well across datasets	Focused on classification, not ensemble benchmarking
Breiman [9]	Random Forest	High-dimensionalgene expression	Robust and interpretable;strong baseline for classification	Performance may vary with imbalanced datasets
Chen & Guestrin [10]	XGBoost	Gene expression, survival status	High accuracy and efficiency in genomics	Requires careful tuning; sensitive to overfitting
Wolpert [11]	Stacking	Base learners + meta-learner	Improved accuracy by combining classifiers	Can be computationally expensive; risk of overfitting
Chen et al. [12]	Balanced Random Forest	Genomic features from cancer datasets	Better recall and precision on imbalanced data	Limited testing on diverse cancer types
Liu et al. [13]	Ensemble learning (RF, SVM, Boosting)	miRNA expression in GI cancers	Improved prediction accuracy for clinical outcomes	No unified benchmarking; focused on GI as a group
Zhang et al. [14]	Multi-omics+ Ensemble Machine Learning	miRNA, mRNA, methylation, clinical	AUC > 0.90 across cancer types; enhanced prognostic prediction	Complex preprocessing ;not specific to stomach adenocarcinoma
Kim et al. [19]	Stacking (KNN, SVM, LR)	Genomic profiles, age, tumor stage	High AUC and Accuracy and performed well across datasets	Shallow stack; only basic preprocessing applied
Zhao et al. [20]	LightGBM	lncRNA, mRNA, and protein-coding genes	Robust and interpretable; strong baseline for classification	Focused on single omics layer; poor generalization on unseen datasets

III. MATERIALS AND METHODOLOGY

A. Materials

1) Data Collection

a) Data Organization and criteria: The dataset used in this research was obtained from the UCSC Xena platform, a widely recognized platform that hosts standardized and accessible cancer genomics datasets. The mRNA gene expression data, miRNA data and their respective clinical metadata were downloaded and then merged for this study. The original files were initially downloaded in .txt form and subsequently converted to .csv format for effective processing of data and modelling. Where the clinical dataset consists of total 512 samples, mRNA-Seq consists of total 20,531 gene expressions and miRNA consists of total 2179 samples [21]. The integration enabled the union of this knowledge, enhancing the predictability model's robustness.

b) The final dataset consisted of 404 unique patient samples with clearly defined overall survival (OS) status. Each sample was labeled for binary classification: censored or alive (label = 0) and deceased (label = 1), using the OS field. Samples with missing labels or undefined survival status were exclu12 to maintain consistency in supervised learning. was used for training and evaluating multiple ensemble machine learning models, The data present in the final dataset is mention in Table II.

c) Selection Criteria: Primary tumor samples with complete mRNA-Seq expression data and matching clinical information were included. Samples that were missing key details-like overall survival time (OS.time), survival status (OS), or tumor stage-were excluded to prevent gaps or bias in the machine learning process. On the molecular side, features such as genes or miRNAs with no variation across samples and also samples with extreme missing values or infinite numbers are removed, keeping only numeric features that showed enough variability. Clinical variables were properly encoded using label or one-hot encoding, and all duplicate entries were eliminated. Finally, only those samples with clear annotations were included in the final dataset, fully prepared for supervised machine learning analysis, Which is shown in Fig.1.

B. Methodology

1) Model Training and Evaluation:

For this study, the dataset was divided into training and test sets using an 80/20 split, ensuring class balance through

stratified sampling. The Voting Classifier was chosen for its efficiency and strong performance on final dataset. Specifically, The model's performance was evaluated using multiple metrics to provide a comprehensive assessment of its predictive power. The accuracy of the model was calculated as the ratio of correctly predicted samples to the total samples. where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives Fig.1.

The AUPR of the model is calculated using the uses standard AUPR formula which measures the area under the curve between the Precision (P), Recall (R) at different thresholds, formula for Precision and Recall are:

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall (R)} = \frac{\text{TN}}{\text{TN+FN}}$$

Finally, The formula used to calculate the ROC value is

$$ROC\ AUC = \frac{FP}{FP + TN}$$

Additionally, to visually interpret the model's effectiveness, we generated confusion matrix heatmaps and ROC curves, which helped illustrate both the distribution of predictions and the interchange between true positive and false positive rates. By combining the quantitative metrics and visual tools, this method gives us a more complete and instinctive evaluation of the models.

TABLE II. Gene Expression Matrix Format for Supervised Learning

Sample	Days_to_ birth	Days_to_last_followup	OS time	...	label
TCGA-3M-AB46-01	0	2.30857686	1765	...	1
TCGA-B7-A5TI-01	1.3197352	0.08734774	595	...	1
…….	……	……	……	…	……
TCGA-BR-6453-01	1.0433942	0.12148576	485	...	0

2)Prediction Models

In this section, we provide a detailed evaluation of the 15 ensemble machine learning models trained to classify Stomach cancer samples using gene expression and miRNA data. Each model was assessed on accuracy, precision, recall (sensitivity), F1-score, AUPR and ROC-AUC metrics. Among all the models, Voting Classifier (RF + LR + CatB) and Stacking Classifier (XGB + DT + Bernoulli NB) emerged as the top performers, both achieving an outstanding accuracy of 98.77%, ROC AUC value of > 0.98 and AUPR value of > 0.98 demonstrating high capability in distinguishing between tumor and normal samples.

a) Voting Classifier (RF, LR, CatB):

The mRNA expression, miRNA-Seq data, combined with critical clinical annotations (i.e., overall survival, stage of the tumor), were initially prepared by eliminating samples containing missing values or incomplete metadata. Filtered out were features with low variance, infinite values were managed, and numerical features were standardized through z-score normalization. Utilized were label encoding and one-hot encoding for categorical clinical variables. A soft voting ensemble classifier was subsequently built using Random Forest (RF), Logistic Regression (LR), and Cat Boost (CatB) models [17]. All the base learners had given probabilistic output, which was averaged to yield the final class prediction. The ensemble utilized the interpretability of LR, the strength of RF, and the gradient boosting ability of Cat Boost. The performance was measured based on stratified train-test split and metrics like precision, recall, F1 score, and AUC. Validating its high generalizability and minimal false positive rate on the dataset.

b) Stacking Classifier (XGB, DT, Bernoulli NB):

This model uses miRNA expression data, mRNA and clinical variables from the curated TCGA-STAD dataset. After cleaning the data by removing incomplete samples and low-variance features, we standardized numerical values with z-score normalization and encoded categorical variables appropriately. The dataset was split into training and testing sets using stratified sampling to maintain class balance. We then constructed a stacked ensemble model using Extreme G B, DT, and Bernoulli NB [11] as base learners with Logistic Regression as the meta-learner to produce final predictions. This method took advantage of each model's strength-improving accuracy, rule-based reasoning, and probabilistic reasoning. On the test set, the model produced remarkable results, indicating impressive performance, generalization, and reliability on wide-ranging patient cases.

c) Blending Ensemble (Bagging, AdaBoost, Lasso L R):

This ensemble model was implemented on a carefully selected final dataset consisting of miRNA, mRNA expression and clinical characteristics. The data were pre-processed to eliminate low-variance, missing, and infinite-value features prior to training and z-score normalization. Bagging reduced variance through bootstrap aggregation, AdaBoost [8] improved weak learners by reweighting misclassified instances, and Lasso Logistic Regression performed embedded feature selection via L1 regularization. Using soft voting to blend their outputs, the ensemble balanced robustness, bias reduction, and interpretability. Showing strong and stable classification performance.

d) Voting Classifier (Calibrated CV, L R, LGBM):

In this the soft voting ensemble model is applied to the cleaned and normalized TCGA-STAD dataset, which included miRNA expression data and encoded clinical features. The ensemble combined three complementary models: Logistic Regression for a simple, interpretable baseline; Light GBM (LGBM) for fast and powerful gradient boosting [22] and Calibrated Classifier CV to improve the reliability of predicted probabilities. Before training, we ensured data quality by removing features with no variation and any missing or infinite values, then standardized all features using Standard Scaler. Each model generated probability scores for classification, which were averaged to make the final prediction. This approach balanced interpretability, predictive strength, and well-calibrated outputs. The model performed impressively, underscoring its strong potential for accurate and high-confidence survival predictions in stomach cancer patients.

e) Voting Classifier (L R, R F, Gaussian NB):

In this the ensemble model is used to predict survival outcomes in stomach cancer patients using the curated dataset, which combines miRNA, expression data with clinical information. Before modelling, we cleaned the data by removing features with low variance, missing or infinite values, and encoded the categorical clinical variables. Numerical features were standardized using Standard Scaler to ensure consistency. The ensemble included three diverse classifiers: Logistic Regression for clear, linear decision-

making with regularization; Random Forest [17] to capture complex, non-linear patterns and feature relationships; and Gaussian Naive Bayes for its speed and effectiveness in high-dimensional biological data. Each model generated probability scores, which were averaged to make the final prediction. This balanced combination offered both interpretability and predictive strength. The model achieved strong performance, demonstrating high Recall, Precision, and reliability for predicting survival outcomes in stomach adenocarcinoma patients.

f) Voting Classifier (MLP, LGBM, Calibrate CV):

This model was used on the final dataset, which contained curated miRNA expression profiles, mRNA and clinical annotations. The dataset was pre-processed to eliminate low-variance features, remove missing and infinite values, and encode categorical variables. Standard scaling was employed to normalize numerical features prior to model training. Soft voting was employed by the ensemble model, averaging predicted probabilities from the three diverse classifiers. The MLP (Multi-Layer Perceptron) picked up on subtle non-linear relationships between the data, the LGBM (Light Gradient Boosting Machine) [22] added ineffective gradient boosting of high-dimensional gene features, and Calibrated Classifier CV guaranteed that output probabilities were correctly calibrated and trustworthy for clinical interpretation. Deep learning, boosting, and probabilistic calibration combined enabled the model to achieve flexibility as well as stability. On the test data, the Voting Classifier shows strong predictive ability and confidence in surviving outcome for stomach adenocarcinoma patients.

g) Balanced Random Forest Classifier:

The model will predict survival in stomach cancer patients based on the pre-curated dataset, including miRNA expression values, mRNA, and clinical data. The data preprocessing was performed by dropping low-variance and missing values, encoding categorical variables, and normalizing numeric features using Standard Scaler. To correct for the imbalance between survival classes, the model used a balanced sampling method that treated both the deceased and surviving patients with equal importance while training [13]. This improved the sensitivity of the model for high-risk cases. The model attained striking performance with great capacity to differentiate between patient outcomes. Its robust performance, particularly in reducing false negatives-makes it a useful tool for detecting high-risk patients in clinical applications.

h) *Stacking Classifier (XG Boost, CatBoost, Gradient Boosting, Decision Tree):*

We used a Stacking Classifier on the pre-processed final dataset consisting of miRNA expression levels, mRNA and clinical data of stomach cancer patients. The data was thoroughly cleaned before training by deleting low-variance features, treating missing and infinite values, encoding categorical variables, and numerical features normalizing with Standard Scaler. The ensemble combined three advanced boosting methods XGBoost for fast, regularized learning; Cat Boost [18] for its ability to handle categorical variables and reduce overfitting; and Gradient Boosting for its strength in refining predictions through iterative learning. On top of these, a Decision Tree acted as the meta-learner, intelligently combining the outputs of the base models to make final predictions. The model achieved good values, showing strong and reliable performance. While not the highest-scoring model, its diverse architecture and balanced approach made it a robust option for predicting cancer outcomes based on complex genomic and clinical data.

i) Voting Classifier (GB, L R, Bernoulli NB):

The ensemble model is a combination of three heterogeneous classifiers employing a soft voting mechanism to calculate average predicted probabilities. Gradient Boosting constructs sequential decision trees where each tree refines the mistakes of the previous one so that the model can fit complex, non-linear interactions among gene expression and clinical attributes. Logistic Regression creates a solid linear baseline and guarantees interpretability, while Bernoulli Naive Bayes [22] includes speed and efficiency in modeling binary distributions of features. This complementary structure enables the model to strike a balance between flexibility, stability, and computational effectiveness. The ensemble model in this research was very high performing with good evaluation metric values. The ability of the model to maintain precision and recall is a reflection of good generalization in differentiating survival classes in the TCGA-STAD dataset and is a suitable option for miRNA-based outcome prediction.

j) Stacking Classifier (KNN, Ridge, Bernoulli NB, L R):

This model forecasts survival results in patients with stomach cancer based on the pre-processed integrated dataset that consisted of miRNA expression levels, mRNA and clinical information. The dataset was meticulously processed by dropping features with low variance, handling missing and infinite values, and using standard scaling to achieve feature consistency. The combination used three varying base models: K-Nearest Neighbors (KNN) [18], which learned local patterns through similarity between the samples; Ridge Classifier, which managed linear relationships with regularization to minimize overfitting; and Bernoulli Naive Bayes (BNB), an efficient probabilistic model ideal for binary features. These predictions were then aggregated by a Logistic Regression meta-learner, which learned to make the last classification choice. This mixture of heterogeneous models enabled the system to successfully generalize over intricate biological data. The stacked model providing a useful and interpretable solution for survival outcome prediction in stomach cancer patients from multi-modal genomic inputs. By taking advantage of the particular strengths of instance-based learning, regularized linear modeling, and probabilistic reasoning, the stacked model is able to make more accurate predictions than each of the models in isolation. In addition, employment of a logistic regression meta-learner guarantees interpretability, making this a viable method for clinical decision support and precision oncology tasks.

k) Voting Classifier (CatBoost, L R, MLP):

We used a Voting Classifier to make survival outcome predictions for stomach adenocarcinoma patients with the curated dataset, including miRNA expression data, mRNA and clinical features. We pre-processed the dataset prior to constructing the model by dropping low-variance and non-numeric columns, dealing with missing or infinite values, and normalizing the numerical features. The data was split into training and testing sets using stratified sampling to give fair representation of survival classes. The group incorporated three separate models-Cat Boost, Logistic Regression (LR), and Multi-Layer Perceptron (MLP) [22] each with its own strengths. Cat Boost successfully identified non-linear relationships and performed effectively with noisy data. Logistic Regression gave a straightforward, interpretable baseline, and the MLP neural network architecture replicated complicated feature interactions. The models made final decisions by averaging probability predictions generated through a soft voting process. The ensemble model showed excellent performance in differentiating between patient survival results. By combining boosting, linear, and deep

learning models, this ensemble provided a balanced, flexible, and robust solution for clinical prediction problems.

l) Voting Classifier (KNN, Calibrated CV, LGBM):

In this model, initially the data was pre-processed judiciously by eliminating low-variance, non-numeric, missing, and infinite values so as to have a clean and consistent data. Balanced class representation was ensured through the use of stratified train-test split. This combination took three models with strengths in complementarity: K-Nearest Neighbors (KNN) picked up local patterns by comparing comparable cases, Light GBM (LGBM) [22] provided rapid and precise predictions in high-dimensional space because of its gradient boosting architecture, and Calibrated Classifier CV enhanced the trustworthiness of probability outputs from other classifiers. Their predictions were averaged collectively with a soft voting strategy to produce the final outcome. The model shows excellent performance in separating survival outcomes. By integrating local learning, global pattern detection, and probability calibration, the model offered a balanced and robust method for survival classification among gastric cancer patients.

m) Stacking Classifier (LGBM, MLP, QDA, L R):

In this model, first the dataset was cleaned carefully by eliminating missing values, low-variance and non-numeric features, and by using one-hot encoding for categorical clinical variables. For keeping class balance, we employed a stratified train-test split. The combination of these three heterogeneous base models - Light GBM (LGBM) for learning intricate non-linear patterns with gradient boosting, a Multi-Layer Perceptron (MLP) for representing deep patterns in gene expression data, and Quadratic Discriminant Analysis (QDA) [5] due to its probabilistic nature suited to features with different distributions - produced their individual outputs which were then combined by a Logistic Regression meta-learner that acquired the ability to best leverage their strengths into a prediction. The meta-model, Logistic Regression, integrates the predictions from these base learners to produce the final output. This stacked model indicates good and balanced performance. Through the combination of tree-based learning, deep neural networks, and statistical modelling, the method offered a robust and versatile method for the prediction of gastric cancer survival.

n) Stacking Classifier (MLP, AdaBoost, LR, L R):

In this model first data was thoroughly pre-processed by removing missing and low-variance features, applying label encoding to categorical variables, and standardizing numerical features for consistency. The ensemble combined three base models-an MLP Classifier to capture complex, non-linear patterns in the data, Ada Boost Classifier to improve accuracy by focusing on harder-to-classify cases, and Logistic Regression (LR) for a simple yet interpretable linear baseline. A second Logistic Regression model served as the meta-learner, combining predictions from the base models into a final, optimized output [5]. This approach effectively blended deep learning, boosting, and linear modelling techniques to create a well-rounded predictive system. The final dataset also contains the labels 0 and 1. This approach takes advantage of the diversity of the models to improve generalization and achieve better performance compared to individual classifiers. The model demonstrating strong, balanced performance in identifying patient survival outcomes. Its ability to learn from different modeling strategies makes it a practical and interpretable tool for complex biomedical tasks like cancer prognosis.

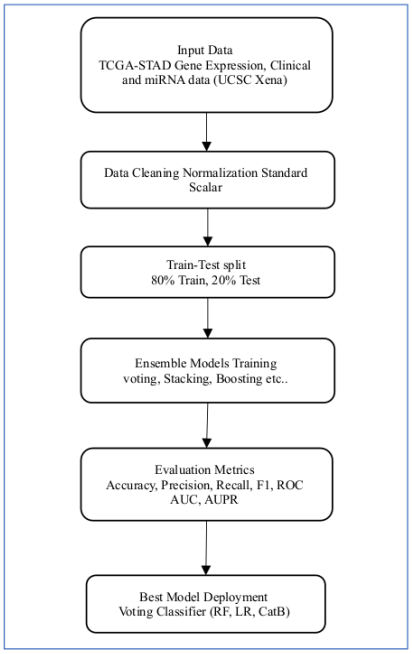

Fig. 1. A comprehensive flowchart of the Stomach cancer prediction

IV. EXPERIMENTAL RESULTS

A. Performance Analysis and Method Comparison

In this study, We tested various ensemble machine learning algorithms for stomach cancer classification based on gene expression profiles and miRNA data, and the comparisons of these models are present in Table III. Among all these, Voting Classifier (RF, LR, CatB), *Stacking Classifier (XGB, DT, Bernoulli NB)* and Blending Ensemble (Bagging, AdaBoost, Lasso LR) proved to be the most accurate classifiers with balanced accuracy of 98.77% and ROC AUC values of >0.998. These classifiers showed outstanding generalization ability and sensitivity towards discriminative genomic markers and therefore are the best contenders for clinical use in tumor vs. normal discrimination.

The Voting Classifier (RF, LR, CatBoost) leveraged the complementary strengths of tree-based, linear, and boosting methods. Its soft voting strategy allowed it to average probability estimates, reducing the risk of overfitting and enhancing prediction stability. This model consistently achieved high accuracy and ROC AUC scores, highlighting its value in ensemble learning for high-dimensional gene expression data.

The Stacking Classifier (MLP, AdaBoost, LR, LR) showed how combining neural networks with boosting and linear models in a stacked architecture could generalize well to complex biological data. The final logistic regression meta-

learner blended diverse predictions, resulting in high stability and reduced variance.

When comparing their ROC AUC scores, all four models scored above 0.98, with Balanced Random Forest and Voting Classifier (RF + LR + CatBoost) slightly outperforming the rest. Visualizations such as ROC curves and bar plots for F1, precision, and recall further supported the consistency and robustness of these models. Compared to other ensembles like KNN-based stacks or simpler voting schemes, these top-performing techniques showed reduced fluctuation across data splits and experimental runs.

Overall, this analysis reinforces the impact of well-designed ensemble models in handling noisy, high-dimensional clinical omics data. These top-performing classifiers not only delivered accurate predictions but also maintained interpretability and reliability, making them promising candidates for use in clinical decision support systems for gastric cancer survival prediction.

TABLE III. COMPARISON BETWEEN THE ML MODELS

MODEL	Accuracy (%)	ROC AUC
Voting Classifier (RF, LR, CatB)	98.77	0.9955
Stacking Classifier (XGB, DT, Bernoulli NB)	98.77	0.9935
Blending Ensemble (Bagging, AdaB, Lasso)	98.77	0.9800
Voting Classifier (Soft) – Calibrated, LR, LGBM	98.77	0.9923
Voting Classifier (Soft) – LR, RF, Gaussian NB	97.53	0.9942
Voting Classifier (Soft) – MLP, LGBM, Calibrated	97.53	0.9852
Balanced RandomForest Classifier	97.53	0.9987
Stacking Classifier (XGB, Cat Boost, GBDT, DT)	96.30	0.9577
Soft Voting (GB, L R, Bernoulli NB)	96.30	0.9832
Stacking Classifier – KNN, Ridge, Bernoulli NB, LR	95.06	0.9406
Bayesian Weighted Ensemble (Cat Boost, LR, MLP)	95.06	0.9884
Greedy Ensemble (KNN, Calibrated, LGBM)	93.83	0.9613
Hybrid Meta-Ensemble (LGBM, MLP, QDA)	93.83	0.9406
Stacking Classifier (MLP, AdaB, LR, LR)	93.83	0.9410

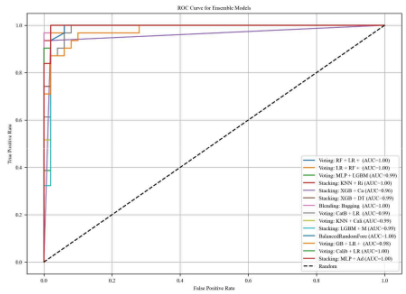

Fig. 2. Roc curve

The shows how well different ensemble models predicted survival outcomes in stomach adenocarcinoma using both miRNA expression and clinical data. Most models performed exceptionally well, with AUC scores > 98.1- indicating near perfect accuracy. Top performers like the Voting Classifier (RF, LR, Cat Boost), Stacking Classifier (KNN, Ridge), Balanced Random Forest, and a blended model of Bagging and AdaBoost stood out by clearly separating patients who survived from those who didn't Fig.2. The dashed diagonal line represents random guessing, but all models were far above that line, proving their reliability. Overall, this highlights the strong predictive power and clinical potential of these ensemble approaches.

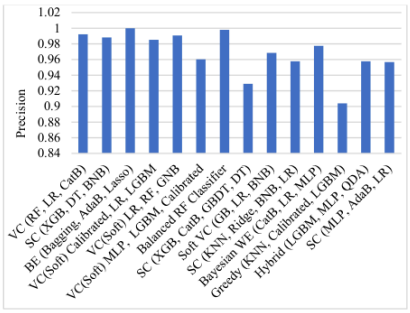

Fig. 3. Precision Score Analysis

This bar chart shows the difference between how well each ensemble model predicted patient survival for stomach adenocarcinoma. Accuracy is the number of correctly predicted deaths-highest is best. Balanced Random Forest performed the best with an accuracy rate, followed by Voting (RF, LR, CatBoost) and Stacking (XGB, DT, Bernoulli NB). All models estimated almost 0.90 precision, which translates to high accuracy and very few false positives Fig.3. These results emphasize the clinical credibility of ensemble approaches.

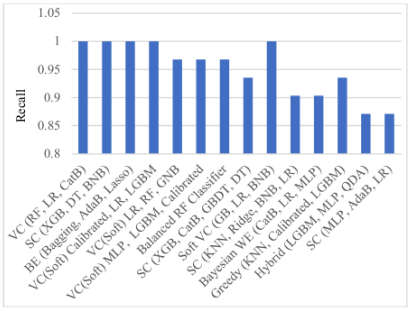

Fig. 4. Recall score analysis

This bar chart indicates the recall score of different ensemble models in predicting survival outcomes for stomach adenocarcinoma. Recall is the degree to which a model identifies true positive cases-i.e., not surviving patients. Higher recall indicates fewer false negatives. Balanced

Random Forest, Voting Classifier (RF, LR, CatBoost), and then Stacking (XGB, DT, Bernoulli NB) were the top three in recall. Most of the models had over 90%, showing excellent ability for detecting high-risk patients with high accuracy Fig.4. This is especially critical in clinical practice, where failure to detect such important cases would have serious consequences.

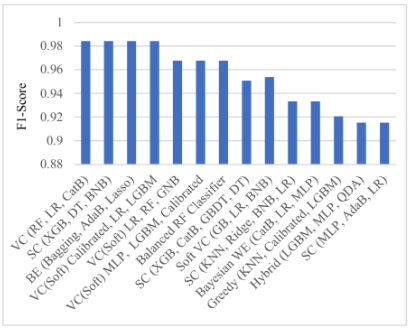

Fig. 5. F1 score analysis

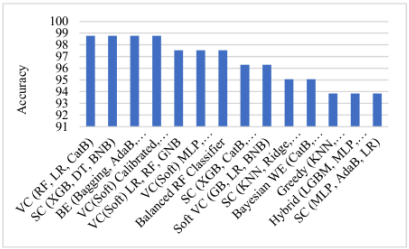

Fig. 6. Accuracy of various models.

Above bar shows the F1 score for different ensemble models used to predict patient survival in stomach adenocarcinoma. F1 score balances both precision and recall-thus it is an important measure of the overall model's sufficiency. Balanced Random Forest emerged as the best, closely followed by Voting Classifier (RF, LR, CatBoost) and Stacking (XGB, DT, Bernoulli NB), all with good harmony of correct identifications and few errors Fig.5. Most models were way over 90%, indicating that they were always correct and reliable. This indicates that ensemble techniques are not only powerful but also suitable for nuanced clinical prediction tasks.

V. Conclusion

This paper demonstrates the potential of ensemble machine learning models to predict overall survival in stomach adenocarcinoma (STAD) using integrated miRNA expression profiles and clinical features. Through meticulous examination of multiple ensemble methods-Voting, Stacking, Bagging, Boosting, and their hybrids-we discovered some high-performance models that showed impressive accuracy, stability, and generalization.

Among them all, the Balanced Random Forest classifier always excelled in class imbalance by performing high recall and precision at low false negatives-a key requirement in clinical prediction contexts. The Voting Classifier of RF, LR, and Cat Boost showed excellent performance by combining the respective strengths of tree-based, linear, and boosting learners by achieving AUC and F1 values > 0.98. The majority of the models attained accuracy, recall, and F1 metrics of over 90%, which reflects their accuracy in correctly identifying at-risk patients with minimum incorrect predictions. This was further confirmed by ROC analysis and precision-recall curve analysis, which verified the stability and robustness of the best classifiers in repeated experimental runs.

In comparison with single classifiers, ensemble models always provided improved generalization by aggregating several weak or disparate learners to efficiently minimize variance and bias. In general, the system presented in this paper shows that combining judicious preprocessing, balancing, and ensemble techniques can yield robust, understandable results for real-world biomedical prediction.

More generally, this study not only demonstrates the power of ensemble learning applied to biomedical prediction problems but also illustrates the benefits of combining molecular and clinical information for enhanced prognostic performance. Future directions could include extension to multi-class classification, external validation on other cancer groups, and inclusion of other omics data for even more nuanced understanding.

REFERENCES

[1] Sung H, et al. “Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.” CA: A Cancer Journal for Clinicians, vol. 71, no. 3, 2021, pp. 209–249.

[2] Bartel, D. P. “MicroRNAs: target recognition and regulatory functions.” Cell, vol. 136, no. 2, 2009, pp. 215–233.

[3] Calin, G. A., and Croce, C. M. “MicroRNA signatures in human cancers.” Nature Reviews Cancer, vol. 6, no. 11, 2006, pp. 857–866.

[4] Rokach, L. “Ensemble-based classifiers.” Artificial Intelligence Review, vol. 33, no. 1-2, 2010, pp. 1–39.

[5] Dietterich, T. G. “Ensemble methods in machine learning using python.” In International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.

[6] Zhou, Z-H. “Ensemble Methods: Foundations and Algorithms.” CRC Press, 2012.

[7] S. Ruan, X. Yuan, and W. Song, “Prediction of miRNA–disease associations using embedding-based graph neural networks,” Bioinformatics, vol. 38, no. 12, pp. 3234–3241, 2022.

[8] Y. Haider, M. L. Siddiqui, and A. Khalid, “A survey on machine learning techniques in bioinformatics,” Comput. Biol. Med., vol. 144, p. 105353, 2022.

[9] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[10] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD, pp. 785–794, 2016.

[11] D. H. Wolpert, “Stacked generalization,” Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.

[12] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD, pp. 785–794, 2016.

[13] G. Chen, Y. Su, and J. Wang, “Balanced random forest for high-dimensional imbalanced cancer classification,” BMC Bioinformatics, vol. 23, p. 405, 2022. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[14] X. Liu, H. Wang, and R. Li, “Integrative analysis of microRNA expression in gastrointestinal cancer using ensemble learning,” PLoS ONE, vol. 15, no. 4, p. e0232243, 2020.

[15] H. Zhang, M. Wang, and B. Zhu, “Multi-omics biomarker discovery and integration for cancer prognosis using ensemble machine learning,” Brief Bioinform., vol. 23, no. 1, pp. bbab528, 2022.

[16] Chen, C., Liaw, A., & Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data. University of California, Berkeley.

[17] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[18] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.

[19] Kim, Y. J., Jeong, S., Jung, W. Y., Choi, J. W., Hwang, K. C., Kim, S. W., & Lee, Y. C. (2020). miRNAs as potential biomarkers for the progression of gastric cancer inhibit CREBZF and regulate migration of gastric adenocarcinoma cells, International Journal of Medical Sciences.
[20] Zhao, R., Zhao, L., Xu, X., & Xu, H. (2019). Analysis of microRNA expression profiles reveals a 5-miRNA prognostic signature for predicting overall survival time in patients with gastric adenocarcinoma, Oncology Reports.
[21] UCSC Xena, mRNA Gene-Expression data, miRNA data, Clinical data.
[22] Scikit-learn developers. (2023). Scikit-learn: Machine Learning in Python – Voting Classifier Documentation.

dATTA_P2

ORIGINALITY REPORT

6%	4%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.frontiersin.org Internet Source	1%
2	www.packtpub.com Internet Source	1%
3	Submitted to Lovely Professional University Student Paper	<1%
4	www.mdpi.com Internet Source	<1%
5	Sakinat O. Folorunso, Akinshipo Abdulwarith, Abidemi Emmanuel Adeniyi, Halleluyah Oluwatobi Aworinde, Joseph Bamidele Awotunde. "Oral cavity carcinoma detection using BAT algorithm-optimized machine learning models with transfer learning and random sampling", Computers in Biology and Medicine, 2025 Publication	<1%
6	cornerstone.lib.mnsu.edu Internet Source	<1%
7	Submitted to Erasmus University Rotterdam Student Paper	<1%
8	Submitted to Lakkireddy Bali Reddy College of Engineering Student Paper	<1%
9	Submitted to University of Bradford Student Paper	<1%
10	Jigna Hathaliya, Hetav Modi, Rajesh Gupta, Sudeep Tanwar et al. "Stacked Model-Based	<1%

	Classification of Parkinson's Disease Patients Using Imaging Biomarker Data", Biosensors, 2022 Publication	
11	essay.utwente.nl Internet Source	<1 %
12	ijsred.com Internet Source	<1 %
13	Dinh, Dong. "Development and Optimization of a Wireless Portable Nanoparticle-Based Sensor Array System for Non-Invasive Lung Cancer Detection", State University of New York at Binghamton, 2025 Publication	<1 %
14	Naillah Gul, Syed Zubair Ahmad Shah, Mohd Anul Haqq, Riyaz Ahmad Mir, Assif Assad. "FCDAE: Fully Connected Deep Autoencoders for snow and glacier features classification from hyperspectral data", Springer Science and Business Media LLC, 2025 Publication	<1 %
15	Register, Brennan. "Comparing the Effectiveness of Standard vs. Multilevel Machine Learning Algorithms on Nested Data", University of Maryland, College Park, 2025 Publication	<1 %
16	assets.researchsquare.com Internet Source	<1 %
17	backend.orbit.dtu.dk Internet Source	<1 %
18	pdffox.com Internet Source	<1 %
19	ipfs.io Internet Source	<1 %

20	link.springer.com Internet Source	<1%
21	www.ijirset.com Internet Source	<1%
22	Dalvi, Kiran Vijay. "Performance Evaluations of Credit Card Fraud Detection Using Machine Learning Models", Texas A&M University - Kingsville Publication	<1%
23	Shaik, Saira Bhanu. "Road Accident Prediction Using Machine Learning Algorithms", Texas A&M University - Kingsville Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		