Cancer Classification Using ML

Michelle Preetham
Department of Biomedical Engineering
Karunya Institute of Technology and Science, Coimbatore, Tamil Nadu, India

Tabitha Kusum Arun
Department of Biomedical Engineering
Karunya Institute of Technology and Science, Coimbatore, Tamil Nadu, India

G. R. Ashisha
Department of Biomedical Engineering
Karunya Institute of Technology and Science, Coimbatore, Tamil Nadu, India

OPEN ACCESS

Manuscript ID:
ASH-2026-130410317

Volume: 13

Issue: 4

Month: April

Year: 2026

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Received: 06.02.2026

Accepted: 10.03.2026

Published Online: 01.04.2026

Citation:
Preetham, M. et al. "Cancer Classification Using ML." *Shanlax International Journal of Arts, Science and Humanities*, vol. 13, no. 4, 2026, pp. 170-75.

DOI:
https://doi.org/10.34293/sijash.v13i4.10317

Abstract

Personalized therapy based on the genetic features of tumors has replaced the one-size-fits-all approach to cancer treatment. This model was developed to facilitate clinical decision-making in individualized cancer treatment by classifying cancer tumors into nine types based on genetics. mutations, this study determined a machine learning-based method for categorizing genetic alterations in cancer. Using a dataset obtained from The Cancer Genome Atlas (TCGA), we assessed four traditional machine learning algorithms (Chen, Yao, & Wang, 2015): Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost. The training and testing results were analyzed and showed a significant difference, with a training accuracy of 87.98 per cent and a test accuracy of 45.60 per cent. Such a strong gap indicates susceptibility to overfitting, particularly with Random Forest and XGBoost, which exhibited the largest train test divergence. Term Frequency-Inverse Document Frequency (TF-IDF) vectorization was used to preprocess the textual clinical data to transform unstructured language into numerical features that may be classified. According to our experimental findings, XGBoost outperformed all other examined classifiers in terms of accuracy, precision, recall, and F1-score while maintaining the lowest log loss. This can be used to treat each type of cancer with a treatment best suited to the type. They can significantly decrease the human workload, as this is currently done manually, while increasing accuracy. In the future, even more classifiers can be used to increase accuracy and provide more reliable and effective treatment.

Keywords: Somatic Mutation Classification, Cancer Genomic Profiling, Clinical Text Analytics, Supervised Machine Learning, Gradient Boosting (XGBoost), TCGA Dataset, Cancer Classification, Machine Learning, XGBoost, Genomic Data Analysis, Precision Oncology.

Introduction

The leading cause of death comprises 10 million deaths annually resulting from cancer, in line with the World Health Organization (Sun, Wang, & Liu, 2019). (Lee. K, Lee. J, & Kim. H, 2019). Typical cancer treatment options often involve standard approaches that may fail to consider the molecular variations in tumors. Additionally, it is very time-consuming and can be inaccurate when performed manually by an oncologist. Current research has formed a classification but has not been implemented anywhere. Thus, this project aims to bridge that gap by automating the process of determining the type of cancer tumor and thus the type of treatment that can be administered specially for that patient. Oncology has seen a revolution owing to precision medicine, which makes it possible to establish treatment plans that can be customized to the genetic makeup associated with each patient's tumor (Lee et al., 2019) & (Kuijjer et al., 2018). Genetic mutations are key elements in the pathogenesis and development of malignancies. Disagreement in mutational patterns can produce heterogeneous reactions to treatment programs; therefore, the importance of

exact classification is to stratify treatment effectively. The Cancer Genome Atlas (TCGA) project has already collected large genomic and clinical archives, which is sufficient to offer unprecedented opportunities for the design of computational methods that can support clinical decision-making (Darmofal, Viale, & Schultz).

Machine learning techniques have shown significant potential in the medical diagnostic and prognostic fields (Zeng, Li, & Zhao). In cancer genomics, these methods enable the identification of latent patterns in high-dimensional complex datasets that might otherwise be hidden when using standard statistical methods. This study focuses on using classical machine learning algorithms to classify genetic mutations using clinical evidence and genomic features. The important contributions of this study include: (i) the evaluation of four machine-learning classifiers in cancer-mutation classification; (ii) the use of TF-IDF vectorization in processing clinical textual evidence; (iii) the performance evaluation using a set of assessment metrics that may be considered relevant in clinical practice; and (iv) the demonstration of the XGBoost dominance in the studied field.

Related Work

Associated Work Machine learning in oncological studies has expanded exponentially over the past few years. Various studies have explored different aspects of cancer categorization using genomic information. Convolutional neural networks form the basis of deep learning paradigms that have been applied to analyze histopathological imaging to aid the detection of cancer (Darmofal, Viale, & Schultz) and have been employed to derive salient information in clinical records and medical texts.

Traditional machine learning approaches remain competitive in the analysis of structured genomics. Ensemble algorithms, especially boosting algorithms, have produced strong results in various biomedical classification problems, which can be explained by their ability to view the effect of complex interactions between features and their ability to survive overfitting [8]. Most studies using TCGA data have focused on gene expression analysis and survival data mining (Kuijjer, Quackenbush, & al, 2018) and (Zhang, Yang, & Xu, 2022). Nevertheless, a combination of clinical textual evidence and genomic features of mutation classification remains an active area of study. This study builds on these preliminary studies by performing a systematic review of classical approaches to machine learning, focusing on clinical applicability and interpretability.

Methodology

Dataset Description

Description of dataset The dataset in this study is anchored on The Cancer Genome Atlas (TCGA), which is a database of cancer genome-wide (Chen, Yao, & Wang, 2015). It includes genetic mutation data concomitant with clinical evidence related to it in the text. Every case is a single genetic variation and documents the gene locus of the mutation on the chromosome, variation-specific mutational change, Clinical Text (expert-annotated clinical records of the consequences of the mutation), and Class Label (categorical name of the mutation). The data will have several classes with different mutation impacts on the therapeutic response. The features are by design clinically relevant because of the focused extraction, thus there are no complex feature-engineering efforts required

Data Preprocessing

The following consecutive steps are involved in the preprocessing pipeline.

Text Cleaning: The text of the clinical evidence is cleaned of unnecessary special characters and whitespace, and reconciled with the consideration of case formatting, thus making the textual data whitewashed before any further analysis. TF-IDF Vectorization: Term frequency-inverse document frequency (TF-IDF) is used to transform textual clinical evidence into a set of numerical feature vectors. This method estimates the significance of lexical items in the clinical corpus relative to each other. The formal definition of TF-IDF transformation is as follows:

TF IDF(t,d)=TF(t,d)×IDF(t)

TF(t, d) is the frequency of term t in document d and IDF (t) is calculated as.

IDF(t)=log(N/df(t))

where N is the total number of documents, and

df(t) is the number of documents that encode Features: Categorical features are numerical encodings of features that are categorical, such as gene names and variant types, with suitable categorical encoding schemes to permit their subsequent ingestion into the downstream models (Cortes & Vapnik, 1995).

Data Splitting: Stratified sampling divides the entire data into training and testing datasets such that the original distribution of classes in the data is preserved in each split and balanced evaluation is also preserved. Classification Algorithms The four commonly used machine-learning algorithms assessed. These four in particular perform well for large datasets, such as TCGA, and are easy to code.

Logistic Regression: This is a linear parametric model that estimates the membership of a category using the logastic function. Although it is a simple step, it provides estimations of parameters that can be interpreted, and it is a fair baseline.

rRandom Forest: This is an ensemble technique, that is, an ensemble of decision trees (trained on resampled subgroups of the data); decisions can be made by majority vote, and the process inherently has the capability to learn non-linear relationships and provide importance measurements on features.

Support Vector Machine (SVM): This is a discriminatory classifier that finds a plane that can maximize the distance between two distinct classes. RBF kernel is employed in order to obtain non-linear decision boundaries.

XGBoost (Extreme Gradient Boosting): This software is used to form decision trees sequentially and is an optimizing gradient-boost software, with each designed to correct the mistakes of the earlier trees. It has regularization mechanisms to reduce overfitting and is known to perform well on structured data tasks.

Evaluation Metrics

The classifiers are evaluated using multiple metrics to provide a comprehensive performance assessment:

Give an overall performance evaluation:

- Accuracy: The percentage of correct predictions of all predictions.
- Precision The fraction of true positives over all positive predictions.
- Recall (Sensitivity): The number of true positives divided by the number of actual positives.
- F1-Score: The harmonic average of recall and precision.
- Log-Loss: Measures the uncertainty of predictions based on probability estimates; lower values indicate better performance

Figure 1 illustrates the overall workflow of our proposed classification framework

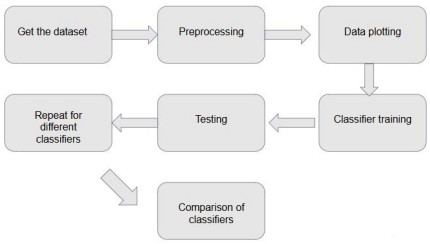

Figure 1 Proposed System Architecture for Genetic Mutation Classification

Experimental Results

Implementation Details

All experiments were conducted using Python 3.x with scikit-learn and XGBoost libraries. The TF-IDF vectorizer was configured with appropriate parameters to balance the vocabulary size and feature informativeness. Hyperparameter tuning was performed using cross-validation to optimize the performance of each classifier.

Performance Comparison

Table 1 Classification Performance Comparison

Classifier	Accuracy	Precision	Recall	F1-Score	Log Loss
Logistic regression	0.6787	0.6722	0.6787	0.6595	1.0155
Random Forest	0.5766	0.5573	0.5766	0.5615	1.3259
SVM	0.4234	0.5129	0.4234	0.3994	1.6425
XCBoost	0.5796	0.5685	0.5796	0.5655	1.2134

Figure 2 shows the confusion matrix for the XGBoost classifier, demonstrating strong classification performance across most mutation classes.

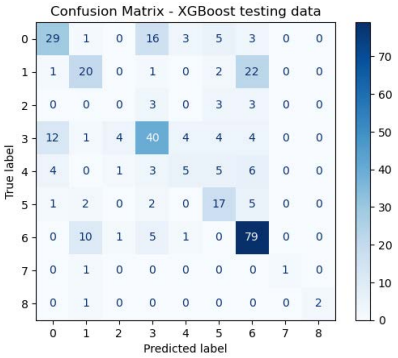

Figure 2 Confusion Matrix for XGBoost Classifier on Test Dataset

Analysis

The experiment showed that XGBoost was always better than the other classifiers for all evaluation measures. This performance is remarkable and can be explained by several factors. XGBoost is based on the Gradient Boosting model that allows decision trees to be built in sequence; thus, the model has a chance to continually rectify its initial mistakes and identify intricate patterns. The addition of L1 and L2 regularization terms is used to prevent overfitting, which is very important in the present case because of the high dimensionality generated by TF-IDF feature derivations. The tree-based structure of the XGBoost algorithm fits the sparse feature matrices produced by TF-IDF vectorization, allowing the model to be trained efficiently. Finally, the fact that XGBoost can capture interactions between features of higher order means that this machine learning model can be used to identify complex associations that may exist between genetic and clinical textual features that would be missed by linear models.

Random Forest is also competitive in nature and enjoys ensemble averaging. However, its performance measures are somewhat lower than those of XGBoost, which could be attributed to the independence of tree construction in the case of the random forest compared to the error-corrective, sequential direction of gradient boosting. Although Random Forest can reasonably work at the baseline, both Support Vector Machines and logistic regression have decision boundaries that are either linear or based on kernels, which can fail to model the complex relationships within the data.

The log loss criterion is useful in clinical settings, where false identifications are severely punishable. The smaller log loss of XGBoost is a feature that shows more precise probability values, which is beneficial in clinical decision-support systems, where the quantification of uncertainty is needed.

Training and testing results were analysed to find that there was a significant difference in training results and test results, where training accuracy was 87.98 per cent and test accuracy was 45.60 per cent. Such a strong gap indicates that there is a susceptibility to overfitting particularly with Random Forest and XGBoost which exhibited the largest train test divergence. The generalization problem is probably due to the complicated problem of categorizing genetic mutations based on very few data and high-dimensional, sparse features.

Limitations and Future Scope

Limitations of the Study

Despite the promising outcomes of the proposed machine learning framework for classifying cancer mutations using both gene mutation and clinical text data, several limitations should be noted.

To begin with, the dataset utilized in this study, based on the Personalized Medicine: Redefining Cancer Treatment dataset, is highly dimensional sparse textual elements produced via TF vectorization using TF-IDF. Although TFIDF is a good method for capturing the relevance of word frequencies, it fails to completely retain the contextual connections between words. Biomedical literature usually includes complicated semantic constructions, and the absence of contextual embeddings can restrict the model's understanding of the biological meaning.

Second, the data have a class imbalance, with some mutation classes having a much larger number of samples than others. This imbalance may favor dominant classes at the expense of minority classes, and hence, lower performance. Despite the fact that stratified splitting was employed to ensure that the distribution of classes in the training, validation, and test sets was preserved, the issue of imbalance still exists and has an adverse effect on generalization performance.

The moderate gap between the training and testing accuracies is another weakness, especially

with ensemble models such as Random Forest and XGBoost. This implies some overfitting based on the complexity of genomic features and small sample size. Cancer mutation data are heterogeneous by definition, and larger datasets may be needed to enhance robustness and stability.

In addition, the analysis was based only on mutation and textual characteristics and did not use other multi-omics data, including gene expression, the most recent methylation patterns, or proteomics. Multimodal biological data would work better in improving prediction and provide a better picture of tumor behavior.

Finally, because the model is computationally feasible in a research context, it would require a significant amount of validation, regulatory approval, and interpretability facilitation to be used safely and ethically in a real-world clinical context.

Future Scope

The findings of this study leave several opportunities to the future study and improvement.

The alternative way out would be to incorporate more sophisticated methods in natural language processing, such as word embeddings (Word2Vec, GloVe) or transformer-based models such as BERT. They can capture a broader range of semantics of the context compared to TF-IDF and can cause a more correct classification of complex biomedical text. The other direction of the future is to take into account deep learning structures, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), to train on the sequential patterns in genetic mutation data. Deep learning Hybrid models Hybrid techniques based on gradient boosting and deep learning can be more effective.

Other integrations, such as multi-omics, the integration of mutation data with gene expression, clinical, and imaging data, should also be considered in future studies. Such integrative models can significantly influence the application of precision oncology, with the option of more precise prediction of tumor subtypes. In addition, to optimize model parameters in a systematic manner and reduce overfitting, hyperparameter optimization techniques, such as Grid Search, Random Search, or Bayesian Optimization, can be used.

Practically, the elaborated model can be extended to a clinical decision support system, which can be used to help oncologists interpret mutations. Applying system XAI models, such as SHAP or LIME, would perhaps allow the system to provide understandable explanations about why the system made the prediction, increasing adoption and confidence in the clinic.

Lastly, the use of artificial intelligence and precision oncology could be employed to accelerate the process of mutation interpretation, reduce the time spent on diagnosing the patient, and plan the personal treatment of cancer patients globally.

Conclusion

The current study provides a comparative study of machine-learning models that are utilized in genetic mutation classification for personalized cancer therapy. The systematic comparison of four classical algorithms, that is, Logistic Regression, Random Forest, Support Vector Machine, and XGBoost, concluded that XGBoost is superior to the others in terms of a range of evaluation measures. TF-IDF vectorization was also found to be effective in converting textual evidence in clinics into numerical form for use in classification.

The poor level of test accuracy of 45.60 per cent as demonstrated across the experiments highlights the complexity of genetic mutation classification as a result of clinical text data. Such a large difference between training and testing performance is an indicator of difficulties in generalization, which has been previously observed in genomics studies with small sample sizes compared to the complexity of features. However, the strengths of XGBoost compared to other competing classifiers prove that the algorithm can be effectively used as a gradient-boosting algorithm to work with high-dimensional and dense attributes based on TF-IDF vectorization.

The results confirm the greater generalizability of machine learning to precision oncology and provide an evaluation framework that can be used in the future to improve the machine learning approach. The comparative framework developed in this study can help researchers and practitioners choose suitable algorithms for use in parallel genetic classification

tasks. Future research will include ways to deal with the identified limitations and achieve greater accuracy. The current model can be used in clinical settings as the first evaluation only and requires mandatory technician analysis. It significantly reduces the workload, although it is not capable of working independently. The current model achieves the objective of classifying the tumour and reducing clinician workload. It greatly improves cancer treatment owing to personalized treatment.

Acknowledgment

We are grateful to say that Karunya institute of technology and sciences gave us the opportunity, infrastructure and academic environment that we still needed to complete this research work successfully. Unremitting motivation and academic assistance that the institution gave me were critical in completing this project.

References

Chen, Y., et al. "Classification of Cancer Primary Sites Using Machine Learning on Somatic Mutations." *BioMed Research International*, vol. 2015, 2015, pp. 1–9.

Cortes, Corinna, and Vladimir Vapnik. "Support-Vector Networks." *Machine Learning*, vol. 20, no. 3, 1995, pp. 273–297.

Darmofal, M., et al. "Deep-Learning Model for Tumor-Type Prediction Using Routine Clinical Sequencing." *Cancer Discovery*, vol. 14, no. 4, pp. 850–861.

Kuijjer, Marieke L., et al. "Cancer Subtype/Type Classification from Somatic Mutations via Pathway De-Sparsification." *PLOS Computational Biology*, vol. 14, no. 4, 2018.

Lee, K., et al. "Accurate Cancer Type Classification Based on Mutation Frequency Patterns (CPEM)." *Scientific Reports*, vol. 9, 2019, pp. 1–11.

Sun, Wang, and Liu. "Identification of 12 Cancer Types through Genome Deep Learning." *Scientific Reports*, 2019, pp. 1–9.

Wang, Liu, R., and H. Zhou. "XGBoost-Based Cancer Classification Using Mutation Frequency Vectors." *IEEE Access*, vol. 8, 2020, pp. 142406–142417.

Zeng, Z., W. Li, and J. Zhao. "Deep Learning for Cancer Type Classification and Driver Gene Discovery from Mutations." vol. 22, no. 1, pp. 1–14.

Zhang, S., L. Yang, and J. Xu. "Multi-Omics Cancer Classification Using Gene Mutations and Expression Profiles." *Frontiers in Genetics*, vol. 13, 2022, pp. 1–12.

Author Details

Michelle Preetham, *Department of Biomedical Engineering, Karunya Institute of Technology and Science, Coimbatore, Tamil Nadu, India,* ***Email ID****: preethammichelle@gmail.com*

Tabitha Kusum Arun, *Department of Biomedical Engineering, Karunya Institute of Technology and Science, Coimbatore, Tamil Nadu, India,* ***Email ID****: tabithaarun66@gmail.com*

G. R. Ashisha, *Department of Biomedical Engineering, Karunya Institute of Technology and Science, Coimbatore, Tamil Nadu, India*