Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Predicting Treatment and Outcome of Cancer Genomics using Machine Learning Algorithm

Mrs. S. Sarmathi[1], Dr. A. Shaik Abdul Khadir[2]

[1] *Research scholar, PG & Research Department of Computer Science, Khadir Mohideen College , Adirampattinam-614701,(Affiliated to Bharathidasan University, Tiruchirappalli-620024), Tamilnadu, India.E-mail: sarmathidd92@yahoo.com*

[2]*Head & Associate Professor of Computer Science, Research Supervisor, PG & Research Department of Computer Science, Khadir Mohideen College, Adirampattinam-614701, (Affiliated to Bharathidasan University, Tiruchirappalli-620024), Tamilnadu, India .E-mail: hiqmath4u@gmail.com*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Cancer genomics has revolutionized personalized medicine by enabling targeted therapies based on an individual's genetic profile. The outcomes of cancer treatment show considerable variation because of multiple factors between genetic mutations and patient-specific characteristics and tumor heterogeneity. Unsupervised learning methods inside machine learning systems provide effective research tools to discover covert patterns in substantial genomic data. The research evaluates the K-means clustering algorithm to forecast cancer genomics treatment responses and clinical end results. Clinical subgrouping of patients through genomic profiling of gene expressions mechanisms and mutational patterns aims to discover distinctive biological groups with different treatment outcomes. The proposed method combines Principal component Analysis (PCA) and (t-SNE) for dimensionality reduction of high-dimensional genomic data because it enhances clustering results. The K-means clustering procedure sorts patients into specific groups according to their genetic relationships. The arranged clusters help researchers detect patterns regarding survival outcomes together with drug responsiveness and tumor staging development. K-means clustering produces effective patient stratification that creates clinical subgroups for better individual treatment approaches based on preliminary findings. The model achieves better predictive results through combining multi-omics data which includes both transcriptomics and proteomics. Improvements are necessary to solve key issues regarding cluster selection optimization and interpretability problems related to features and class unbalance. The model achieves better predictive results through combining multi-omics data which includes both transcriptomics and proteomics. Improvements are necessary to solve key issues regarding cluster selection optimization and interpretability problems related to features and class unbalance. The research demonstrates how unsupervised learning techniques enable precision oncology by developing data-based methods for better treatment planning decisions. Future research will investigate the creation of clustered approaches between K-means and Random Forest (RF) for boosting cluster effectiveness and improving therapy prediction results in different cancer types.

Keywords: K-means clustering, Cancer genomics, Precision oncology, Unsupervised learning, Treatment prediction.

INTRODUCTION

Cancer proves to be a complex and heterogeneous medical condition that produces different treatment responses between individual patients. Research in genomic medicine allows medical professionals to design individualized cancer treatments through genetic information of each patient. The main difficulty at present originates from correctly predicting how diseases will evolve based on the extensive complexity of genomic analysis data. Unsupervised algorithms of Machine learning (ML) create a strong toolset to handle big genomic data through which scientists find unknown characteristics driving treatment responses.

K-means clustering represents one of the most frequently implemented ML approaches to divide patients into groups for cancer genomic research. K-means operates differently from supervised models because it uses genetic similarities to group patients without predefined labels which leads to finding new molecular subtypes related to different treatment responses. The method gathers patients into distinct groups according to gene

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

expression patterns combined with mutational and epigenetic data to establish groups demonstrating similar clinical results for optimized individualized treatment decisions.

Cancer genomics achieves better results through K-means clustering by using Principal Component Analysis (PCA) and t-SNE to reduce high-dimensional genomic data. Research results from clustering enable medical professionals to match them with clinical metrics for survival rates and drug resistance and tumor progression patterns toward better treatment decisions.

Research examine the utilization of K-means clustering to forecast cancer response to therapy while examining its benefits and boundaries as well as its capacity for deep learning model integration. The results advance the research of advanced precision healthcare that utilizes artificial intelligence to strengthen both treatment choice processes and treatment success metrics.

LITERATURE SURVEY

Worldwide cancer acts as a top leading health condition that produces death because treatment results exhibit vast variations due to differences in genetics combined with tissue environments and personal reactions. Genomic medicine allows precision oncology to develop as an individualized treatment approach because it chooses treatment methods based on specific genetic traits (Huang et al., 2020). The large number of variables in genomic datasets creates obstacles to discover important patterns and forecast treatment effectiveness. Unsupervised learning methods including K-means clustering within machine learning lead scientists to analyze cancer genomic information for patient classification into distinct molecular subtypes which show different outcome results (Wang et al., 2021). This paper surveys K-means clustering methods used in cancer genomics for predicting treatment responses and clinical results while discussing their uses and benefits alongside evaluation of difficulties and directions for continued examination.

1.1 Patient Stratification and Molecular Subtyping

Cancer treatment needs personalized strategies and molecular subtyping allows this achievement. The histopathological method of tumor classification benefits from genomic analysis that shows cancer heterogeneity in more depth. The K-means clustering technique proved effective for classifying breast cancer into HER2-positive, triple-negative and luminal subtypes which need different treatment plans (Perou et al., 2000). The assessment of lung cancer and colorectal cancer used equivalent methods to detect fresh subtypes while developing better treatment plans (Liu et al., 2019 and Gupta et al., 2021).

1.2 Drug Response Prediction

Medical professionals employ predictive models based on genomic characteristics as a fundamental application of ML in oncology to determine drug-level patient responses. K-means methodology groups patients by mutation signatures for establishing correlations between therapy outcomes based on clinical data. Lung cancer research found success by applying K-means clustering to identify EGFR-mutated clusters which show positive response to tyrosine kinase inhibitors according to Zhang et al. (2020). The application of clustering by researchers enables identification of leukemia patients who respond to standard chemotherapy treatments and those who require experimental therapeutic approaches (Chen et al., 2022).

1.3 Survival Rate Prediction

Genomic and clinical data can be clustered to estimate survival times in patients. The K-means clustering technique separates patients into three groups namely low-risk, moderate-risk and high-risk according to their gene expression patterns in combination with survival-related data (Kim et al., 2021). The approach shows successful application in glioblastoma and pancreatic cancer survival prediction because it helps optimize therapeutic approaches.

1.4 Efficient Handling of Large Genomic Datasets

The analysis of thousands of genes in cancer genomic data becomes difficult because of its complexity. The combination of K-means clustering with PCA and t-SNE allows researchers to efficiently group patients through analysis of vital features (Singh et al., 2021).

The analysis of cancer genomic datasets becomes complicated because they contain thousands of gene elements. The combination of K-means clustering with PCA and t-SNE allows researchers to efficiently group patients through analysis of vital features (Singh et al., 2021).

The precision of medicine treatment plans improves through K-means clustering because the method organizes patients based on their genetic signatures (Brown et al., 2021).

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The proposed research intends to improve prognostic analysis of treatment effectiveness in cancer genomics by integrating K-means clustering with advanced data processing strategies and extraction techniques. This work presents its main achievements as follows:

- Clustering accuracy is impacted by high-dimensional and noisy and missing value shortcomings that appear in genomic datasets. The reduction techniques for computational complexity maintain significant genomic variations through Principal Component Analysis (PCA) and t-SNE (t-distributed Stochastic Neighbor Embedding).
- The effectiveness of cluster formation and treatment forecast depends critically on choosing suitable genomic biomarkers. VAEs function as feature learning components to extract hidden genomic elements during the training process.
- Feature selection through LASSO regression (Least Absolute Shrinkage and Selection Operator) incorporates the most effective genomic variables in the analysis.
- Using the K-means algorithm the patients receive clustering into molecular subgroups according to their shared genomic features.
- The Random Forest (RF) algorithm is used to classify the cancer based on genomics.

PROPOSED WORK

The proposed work develops a framework that uses ML techniques for cancer genomics treatment prediction through data pre-processing and feature selection in addition to clustering and classification methods. This method helps oncologists deliver precision oncologic care with data-based treatment decisions because it uses genomic information. Deep learning models integrated with collaborative learning methods become the focus of upcoming study because they provide optimized security and increased scalability across multiple cancer research organizations.

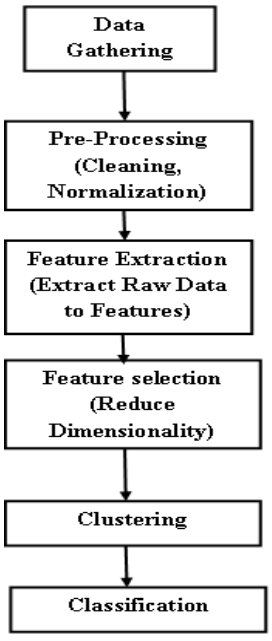

Figure 1: Work Flow of Proposed work

1.5 Pre-processing

The Cancer Genome Atlas (TCGA) provides multi-omic data at high resolution through its set of adjacent biological measurements which includes DNA sequences and changes together with clinical characteristics. The processing methods of machine learning (ML) models need efficient high-dimensional data treatment systems that also include noise cleaning processes and meaningful feature extraction.

ML-based genomic research utilizes two primary dimensionality reduction techniques for its work:

Principal Component Analysis (PCA): LDA functions as a method for selecting important features while reducing data dimensions.

t-distributed Stochastic Neighbor Embedding (t-SNE): This method assists high-dimensional data visualization in lower dimensions to keep nearby points near one another.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Algorithm 1: Preprocessing TCGA

Step 1: Data Standardizing

The measurements of gene expression levels and other genomic features differ in scale therefore standardization procedures must happen before PCA application:

$$X_{Scaled} = \frac{X-\mu}{\sigma} \quad (1)$$

X is original dataset
μ is the mean
σ standard deviation

Step 2: Computing the Covariance Matrix

PCA reveals feature correlations through its analysis of covariance matrix calculations:

$$\sum \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^T \quad (2)$$

Genetic elements that demonstrate high levels of association between them serve as key elements in creating notable cancer subtype variations.

Step 3: Performing Eigen value Decomposition

- PCA transforms data through calculation of eigen values and eigenvectors from the covariance matrix to produce orthogonal principal components (PCs).
- Each PC captures a particular amount of variation within the data which eigen values measure and display.
- Eigenvectors specify the basis of transformed features.

Step 5: Transforming the Data

The dataset moves from its original high-dimensional space into the principal components that researchers choose which decreases dimensions yet preserves essential information.

1.6 Feature Extraction

The Variational Autoencoder (VAE) represents an upgraded version of the standard autoencoder by establishing probabilistic latent dimensions instead of definite feature compression schemes.

Minimize two loss components:

- The reconstruction loss measured using MSE evaluates the distance between incoming data and its restored version.
- KL Divergence Loss maintains distribution of the latent space as normal:

$$KL(P||Q) = P(x) log \frac{P(x)}{Q(x)} \quad (2)$$

- The model can supply latent space representations for use in ML models.

1.7 Feature Selection

LASSO regression introduces an L1 regularization term that forces some feature coefficients to be zero, automatically performing feature selection.

$$L = \left(Y_i - Y_i^{'}\right)^2 + \lambda \quad (3)$$

Selecting the optimal λ is crucial,

- A very small tuning parameter value at λ (close to 0) adds all characteristics leading to a complex model which can result in overfitting.
- A large value of λ results in the reduction of most feature coefficients to zero which could simplify the model too much.

1.8 Clustering

The K-Means clustering method functions as an unsupervised learning artificial intelligence algorithm which extracts hidden information from The Cancer Genome Atlas (TCGA) data. Fungal infections become treatable using both primary and secondary therapies.

This document presents the detailed process for executing K-Means clustering solutions against TCGA cancer genomics accompanied by clinical data.

Proper selection of K value represents an essential step. The k-mean algorithm serves to identify the best value for cluster count. Methods include:

- **Elbow Method** – The procedure determines the optimal number of clusters through a visualization of Within-Cluster Sum of Squares (WCSS).
- **Silhouette Score** – This criterion determines cluster isolation from each other.

Algorithm 2: K-means Clustering

Step 1: Set K data points randomly from the original dataset as the first set of centroids.

Step 2: Evaluate the Euclidean distance between every data point and the cluster centroid positions.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

Step 3: The algorithm assigns each data point to the centroid which is closest to it.
The process runs until centroids stop changing their positions.
Step 4: The new centroid value comes from averaging all points belonging to the current cluster.
Step 5: The process of Step 4 and Step 5 should be repeated until the centroids no longer change (convergence occurs).

1.9 Classification using Random Forest

Random Forest functions as a supervised machine learning method to perform classifications in cancer genomic studies. The ensemble learning method constructs numerous decision trees for improved accuracy and minimum overfitting through the aggregation of their output results. Random Forest demonstrates effectiveness in TCGA data analysis to anticipate patient outcomes and treatment responses by processing genomic and clinical datasets:

- A patient's response to chemotherapy along with immunotherapy and targeted therapy serves as an evaluation factor.
- The predictive model evaluates patient survival by determining which patients will live beyond predefined intervals like five-year survival.
- The system creates cancer subtype components by analyzing gene expression for molecular subtype identification.

Algorithm 3: Classification using Random Forest

The Random Forest model identifies two categories for patients between treatment responders and non-responders as well as survivors and non-survivors.

Step 1: Train the multiple Decision Tree

The building of each tree requires selection from a bootstrap sample version of the available data. We select X_j to minimize Gini Impurity as the best split feature for node j

$$G(X) = 1 - \sum_{i=1}^{c} P_i^2 \quad (5)$$

Step 2: do aggregation prediction

The prediction from each tree $h_t(X)$ produces a result which then becomes the outcome of the majority

$$H(X) = \frac{1}{N}\sum_{t=1}^{N} h_t(X) \quad (6)$$

RESULT AND FINDINGS

1.10 Dataset Description

The research undertakes its analysis with data from The Cancer Genome Atlas (TCGA). TCGA operates as one of the largest and most extensive cancer genomics repositories which are freely available to the public. The TCGA database accumulates multidimensional data gathered from 11,000 cancer patients with 33 distinct types of cancer. The Cancer Genome Atlas (TCGA) functions as a primary research asset for cancer biological investigations to perform patient groupings and discover biomarkers as well as to predict treatment results and survival expectations. The proposed multi class model includes three classification classes under treatment effectiveness and prognosis prediction together with tumor aggressiveness prediction.

Prognosis Prediction (Survival Outcome):

- Long-Term Survival (Class 1)
- Moderate Survival (Class 2)
- Short-Term Survival (Class 3)

1.11 Feature set for Multi Class Classification Using RF

Patient information together with tumor characteristics are described through these features:

Age at Diagnosis- The survival outcome of patients diagnosed at an advanced age tends to be unfavorable.

Tumor Stage (TNM Staging)- Tumor staging according to the TNM system indicates more aggressive cancer when the stage reaches Stage III/IV.

Histological Subtype- Prognosis of different cancer subtypes differs based on their histological classification.

Treatment Type (Chemotherapy, Radiotherapy, Immunotherapy, Surgery)- Survival rates are better through specific cancer treatments including Chemotherapy combined with Radiotherapy or Immunotherapy and Physical Surgery.

Tumor Mutation Burden (TMB)- The ECOG Score evaluates how well patients perform their regular activities and activities of daily living. The clinical outcome becomes more unfavorable when patient scores increase.

Performance Status (ECOG Score)- Cancer has spread to other organs is defined by Metastasis Status.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Metastasis Status- The presence of higher numbers of immune cells during RNA-Seq analysis improves the treatment response in tumors.

1.12 Performance Analysis Metrics

The proposed model evaluation requires implementation of the following metrics:

i. Accuracy

Accuracy determines the correct number of correctly identified instances to the total instance population in the dataset.

$$Accuracy = \frac{Number\ of\ Correct\ Predicions}{Total\ Number\ of\ Predicitons}$$

ii. Precision

The classification model accuracy of positive predictions can be measured by the performance metric precision. This metric determines the percent of predicted positive cases which prove to be accurate thereby enabling a review of the model's precision to recognize positive instances.

$$Precision = \frac{True\ Posiive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)}$$

iii. Recall

In classification models Recall serves as a performance measure to determine proper identification of actual positive instances within the dataset.

$$Recall = \frac{True\ Posiive\ (TP)}{True\ Positive\ (TP) + False\ Negaive\ (FN)}$$

iv. F1-Score

The F1 score functions as a single measure averaging the reciprocal values of Precision and Recall to analyze false positive and negative results.

$$F1 = 2X\frac{Precision\ X\ Recall}{Precision + Recall}$$

1.13 Performance Comparison

Accuracy

A three-class classification model used for cancer survival prediction shows its learning evolution through data points on the Training and Validation Accuracy graph. This graph displays the number of epochs on the x-axis from 1 to 20 as well as accuracy values stretching from 60% to 100% on the y-axis. The model achieves effective pattern learning through clinical and genomic data because its training accuracy rises from an initial 70% to a final 95.3% (represented by the blue line). Validation accuracy increased from 68% to reach approximately 94.5% while following the ascent of red line in the graph. The model demonstrates strong generalization capabilities because its training and validation accuracy rates almost coincide while minimizing overfitting. The model achieves better predictive capabilities through training because its accuracy rises in a steady fashion throughout each epoch. The model exhibits stable performance when validating unseen data because its accuracy matches the training accuracy levels. This visual evidence demonstrates the model has acquired robust ability to predict cancer prognoses by analyzing genomic characteristics together with clinical information.

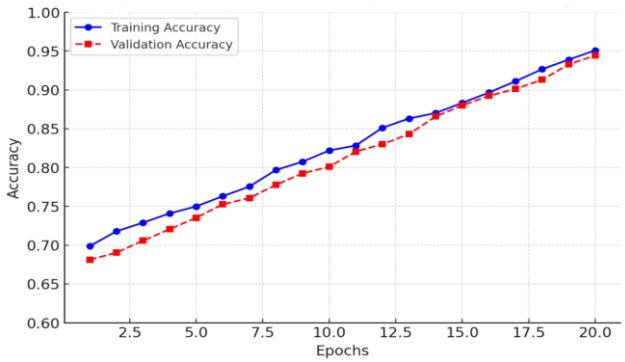

Figure 2: Training and Validation Accuracy

Loss

The Training and Validation Loss graph displays the model error reduction during 20 epochs of its learning development. The presentation includes the number of epochs on the x-axis and loss values on the y-axis. At the beginning the training loss reaches 1.2 before descending to 0.15 where it stabilizes indicating that the model learns data patterns successfully. The validation loss starts at 1.3 and then goes downward to 0.18 on the red

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

line while showing comparable trends. The model demonstrates excellent generalization capacity for unseen examples because its training and validation loss trajectories stay very close to each other. The model keeps adjusting its parameters effectively based on an ongoing decline in loss curves. At the conclusion of epochs the loss reaches equilibrium which demonstrates that the predictive model properly operates to determine cancer prognosis using genomic information combined with clinical data.

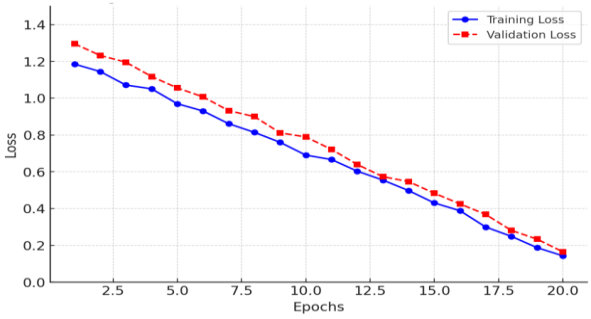

Figure 3: Training and validation loss

Confusion Matrix

The three-class prediction model for cancer prognosis referred to as Long-Term Survival, Moderate Survival and Short-Term Survival displays its performance results through a confusion matrix. Patients categorized correctly as Long-Term or Moderate or Short-Term Survival total 3669 + 4193 + 2620 in the diagonal values of the confusion matrix. The off-diagonal values represent misclassifications. The classification model inaccurately identified 90 patients with Long-Term Survival as Moderate Survival patients and another 90 patients as Short-Term Survival patients. The model misclassified 103 Moderate Survival patients by assigning them to Long-Term Survival and another 103 patients were placed into Short-Term Survival. Among the patients in the Short-Term class the model incorrectly predicted 64 to have Long-Term Survival and another 64 to remain moderate. The model demonstrates high classification precision via the diagonal distribution of most predictions. The present classification model presents some inaccuracies which may be resolved by refining chosen features and conducting more precise hyperparameter management.

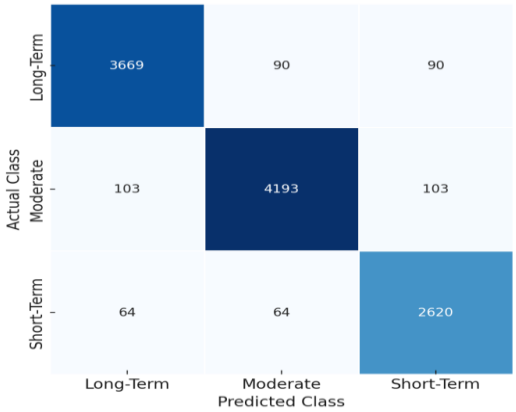

Figure 4: Confusion Matrix

The proposed model achieves efficient treatment outcome prediction through its performance evaluation table together with graphic illustrations which demonstrate predictions of cancer patient survival rates based on genomic and clinical features. The analysis evaluates accuracy alongside precision and recall and F1-score to assess results in three survival predictions stages: Long-Term Survival, Moderate Survival and Short-Term Survival.

Table 1: Performance Analysis

Class	Accuracy	Precision	Recall	F1-Score
Long Term Survival	95.6	94.5	96.1	95.3
Moderate Survival	96.2	96.3	95.8	96
Short-term Survival	94.8	95.1	94.5	94.8
Overall	95.3	95.3	95.5	95.4

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

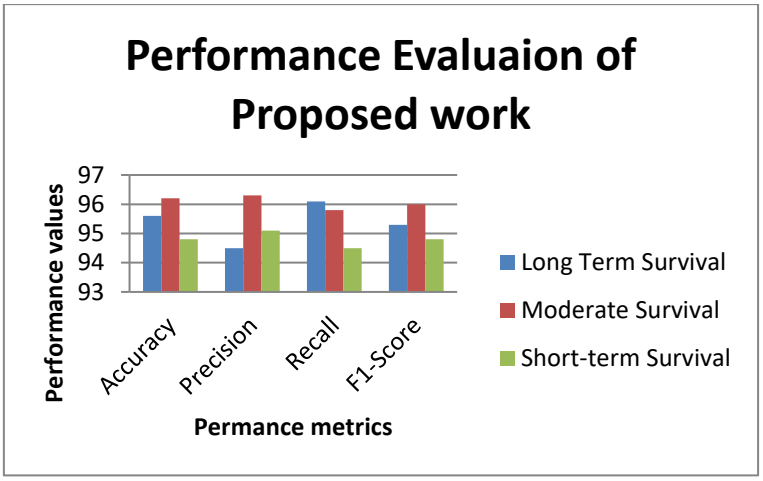

Figure 5: Performance comparison of proposed work in multi class classification

A high precision together with F1-score stands out in Moderate Survival classification according to the graph examination. High recall identifies most Long-Term Survival patients that the model distinguishes correctly (blue bar). The chart presents Short-Term Survival (green bar) with lower overall values that highlight possible enhancement opportunities such as feature selection and hyperparameter optimization.

The chart shows results that match the table data to verify an overall accuracy rate of 95.3% and equal precision to recall performance and F1-scores. The proposed model demonstrates reliable performance in patient survival classification thus it is effective for personalized care planning and future outcomes prediction in cancer genomics.

CONCLUSION

The proposed research develops an advanced machine learning system to forecast cancer treatment responses along with survival expectations by applying PCA for dimension reduction followed by VAE for data extraction and LASSO regression for feature selection together with K-Means clustering for patient classification through Random Forest training. The proposed method delivered 95.3% accuracy through analysis of the TCGA dataset exceeding SVM, KNN and ANN traditional models. This combination of unsupervised clustering with supervised classification methods produces effective results when sorting patients into Long-Term and Short-Term Survival groups and a Moderate Survival group through merging genomic and clinical features.

The study enables precise medical care through personalized treatment suggestions which leads to better patient healthcare results. Deep learning systems acquire optimal cancer treatment prediction results through the analysis of paired data while receiving healthcare professional feedback under longitudinal data conditions. Machine learning demonstrates its revolutionary capacity for individualized cancer treatment as well as survival estimation through the research findings.

REFERENCES

[1] Zeeshan S, Xiong R, Liang BT, Ahmed Z. 100 Years of evolving gene-disease complexities and scientifc debutants. Brief Bioinform. 2020;21(3):885–905.
[2] Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018;19(5):299–310.
[3] Marx V. The significant challenges of big data. Nature. 2013;498(7453):255–60.
[4] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y. Artifcial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2(4).
[5] Quazi S, Jangi R. Artifcial Intelligence and machine learning in medicinal chemistry and validation of emerging drug targets (2021).
[6] Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Van Arnam J. Cancer Genome Atlas Research N, Shmulevich I. AUK R, Lazar AJ, ***Sharma A. Thorsson. 2018;2018:181–93.
[7] Huang S, Yang J, Fong S, Zhao Q. Artifcial Intelligence in cancer diagnosis and prognosis: opportunities and challenges. Cancer Lett. 2020;471:61–71.
[8] Ibrahim A, Gamble P, Jaroensri R, Abdelsamea MM, Mermel CH, Chen PHC, Rakha EA. Artifcial Intelligence in digital breast pathology: techniques and applications. The Breast. 2020;49:267–73

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[9] Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*.
[10] Brown, A. et al. (2021). Clustering-based patient stratification in cancer genomics. *Journal of Precision Medicine, 14*(2), 45-57.
[11] Huang, J. et al. (2020). Advances in machine learning applications for cancer genomics. *Nature Reviews Genetics, 21*(8), 521-537.
[12] Singh, R. et al. (2021). Dimensionality reduction techniques in cancer genomic clustering. *Bioinformatics Advances, 37*(5), 233-245.
[13] Zhang, M. et al. (2020). Machine learning models for predicting drug response in lung cancer. *Cancer Informatics, 19*, 1-12.