Analyzing Cancer Prognosis with Advanced Machine Learning Models

Dr. G. Prabhakar Raju, Maradapu Ananya Sreshta, Kadaverugu Sai Aishiu Preetham, Pisati Bhanuprakash Reddy, Sarabudla Harshitha,
Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana

Abstract
This study investigates the use of Support Vector Machines (SVM) and other machine learning algorithms for predicting the prognosis of lung, breast, and cervical cancer patients. The research evaluates the predictive accuracy, influential features, algorithm performance, and model generalization across different cancer types. Using publicly available datasets, the study highlights SVM's exceptional accuracy in prognosis prediction, with findings indicating its superiority over alternative algorithms. Notably, it identifies the key clinical, molecular, and pathological features that significantly impact predictive accuracy. The study also discusses the clinical applicability of these models, emphasizing their potential to aid healthcare professionals in making more informed treatment decisions. Acknowledging limitations, including data availability and computational resources, the study suggests future directions, encouraging the exploration of additional techniques, diverse datasets, and real-world clinical trials to validate the model's effectiveness.

Keywords
Cancer, Medical Diagnosis, Markers, Learning, Patients, Machine Learning, Support Vector Machine

Introduction

Cancer, especially lung, breast, and cervical cancers, is a major health challenge. Knowing how serious a patient's cancer is and what the best treatment might be is critical. We can use computer programs, like Support Vector Machines (SVM), to help with this. In our study, we want to see how good SVM and other similar programs are at predicting what might happen to cancer patients. We will look at how accurate these programs are, which factors are most important, how well they perform compared to others, and if doctors can use them to help patients. By using a lot of information available to everyone, we hope to make cancer predictions better, so doctors can give patients better care.

This study is not just about computers; it is about how we can make medical care better. We will not only see if these computer programs work well but also if doctors can use them to help make decisions. We know there are some limits to our study, like not having all the data we want, but we will also talk about what we can do in the future to make this even better. So, let us explore how computers can help make cancer treatment smarter and more effective.

Literature Survey

Kwetishe Joro Danjuma presents a study conducted at Modibbo Adama University of Technology, Nigeria. The research focuses on assessing the effectiveness of various machine-learning algorithms for predicting post-operative life expectancy in lung cancer patients. The study involves the Department of Computer Science and highlights the importance of accurate prognosis in improving patient care. The article contributes to the field of medical decision-making by exploring the potential of machine-learning techniques in the context of lung cancer treatment outcomes [1].

Mandhir Kaur and Dr. Rinkesh Mittal proposed a comprehensive overview of intelligent techniques employed for brain tumour detection. The study delves into computer science and medical imaging to explore advancements in detecting brain tumours. Through this survey, the authors address the challenges and opportunities in this domain, highlighting the significance of accurate and efficient diagnosis. The article contributes to the field by synthesizing various intelligent methods and their applications. It is a valuable resource for researchers and practitioners working in brain tumour detection and medical imaging. A comprehensive overview of various intelligent methods employed for detecting brain tumours. The survey explores and analyses a range of techniques such as image processing, machine learning, and artificial intelligence, which play a vital role in enhancing the accuracy and efficiency of brain tumour detection. By examining the advancements and challenges in this field, the authors contribute to the understanding of state-of-the-art methodologies and their potential applications in diagnosing brain tumours. This survey is a valuable resource for researchers, practitioners, and medical professionals working on improving brain tumour detection and treatment strategies [2].

Zehra Karhan and Taner Tunç present a study on lung cancer detection and classification using various classification algorithms. The research explores the application of these algorithms to distinguish between lung cancer cases, aiming to enhance diagnostic accuracy. Through their investigation, the authors contribute insights into the effectiveness of classification techniques in the context of lung cancer diagnosis. This paper serves as a valuable resource for researchers and practitioners in medical imaging and cancer detection, offering a comprehensive examination of classification methods for lung cancer identification [3].

Ada and Rajneet Kaur, published in the International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013, presents a comprehensive study on the application of data mining classification techniques for the detection of lung cancer. The authors delve into these techniques to enhance the accuracy and efficiency of lung cancer detection, which is crucial for early diagnosis and treatment planning. Through their research, Ada and Rajneet Kaur contribute valuable insights into the integration of data mining approaches in medical diagnosis, particularly focusing on lung cancer detection. This study serves as a significant resource for researchers and practitioners in the field of medical data analysis and cancer detection, offering an in-depth exploration of data mining techniques' potential in improving lung cancer diagnosis [4].

The primary objective of the paper is to explore the use of an extreme learning machine (ELM) for the multicategory classification of different cancer types based on microarray gene expression data given by R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran [5].

Y. Xiao, J. Wu, Z. Lin, and X. Zhao introduce an innovative approach to cancer prediction by combining multiple deep-learning models through an ensemble technique. The findings suggest that this multi-model ensemble approach can enhance cancer prediction accuracy by leveraging the diversity of deep learning models [6].

Summary of Literature

- *One study uses computer algorithms to predict how long lung cancer patients will live after surgery.*
- *Another study explores ways to find brain tumours using computer techniques, helping doctors diagnose them accurately.*
- *There is research on using different computer methods to identify lung cancer cases with high accuracy.*
- *A paper discusses using data mining techniques to improve the detection of lung cancer, helping with early diagnosis and treatment.*
- *One study uses a computer model called an "extreme learning machine" to classify different cancer types based on gene data.*
- *Lastly, a group of researchers found that combining multiple advanced computer models can make cancer predictions more accurate.*

Methodology

Architecture

The following figure shows us the flow chart of the proposed methodology:

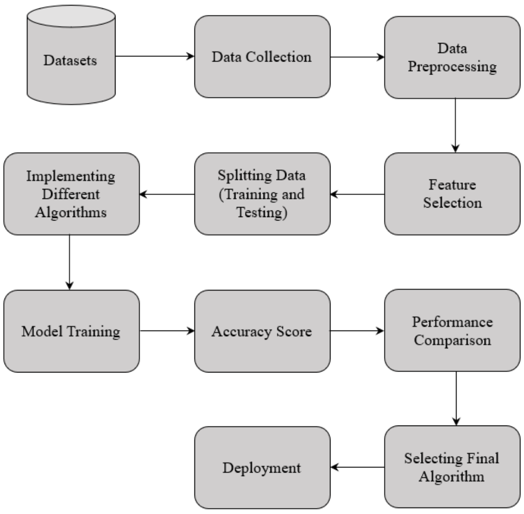

Figure: 1 Architecture of Proposed Methodology

Data Collection

The datasets collected from the UCI Machine Learning Repository and data world, these datasets contain clinical, molecular, behavioural and environmental features influencing cancer patients.

Data Preprocessing

In this stage, we prepare the data for further refinement by adjusting, removing, and optimizing it to meet specific requirements. We begin by scrutinizing the data types to ensure they are appropriate. Null values are identified and addressed by replacing missing categorical values with the mode (the most frequently occurring value) and continuous values with the mean. We then assess the distribution of the target variable to identify any data imbalances, where one class is disproportionately represented. To facilitate modelling, we encode the target variable, converting categorical values into numerical representations. This phase ensures that the data is well-prepared for subsequent processing and analysis.

1. **Check Data Types:**
   - Verify the data types of each feature to ensure they are appropriate for further processing.
   - For example, numerical features should be represented as floats or integers, while categorical features should be represented as strings or integers.
2. **Handle Null Values:**
   - Check for missing or null values in the dataset. This can be done using functions like ".isnull()" or ".isna()" in pandas.
   - For categorical features, replace null values with the mode (most frequent value) of the respective feature.
   - For continuous features, replace null values with the mean or median of the respective feature.
3. **Check for Data Imbalance:**
   - Examine the distribution of the target variable to check for data imbalance. Data imbalance occurs when one class of the target variable is significantly more prevalent than others.
   - This can be visualized using histograms, bar plots, or value counts.
   - If an imbalance exists, consider techniques such as oversampling, under-sampling, or using algorithms that handle class imbalance (e.g., weighted classes).

4. **Encode Target Variable:**
   - If the target variable is categorical, encode it into numerical values suitable for modelling.
   - For binary classification, you can use label encoding (e.g., replacing 'Yes' and 'No' with 1 and 0).

Feature Selection

In this process, the dataset is further processed to choose the required features and eliminate the features not needed. The following tables show basic features in different cancer datasets:

Features	Meaning
Patient Id	*Patient ID details*
Age	*Age of Patient*
Gender	*Gender of Patient*
Air Pollution	*Air Pollution in the Patient Environment*
Alcohol use	*Alcohol Consumption of the Patient*
Dust Allergy	*Dust Allergy, if any of the Patient*
Occupational Hazards	*Occupational Hazards Prone to the Patient*
Genetic Risk	*Genetic Risk of Lung Cancer Occurrence in the Patient*
Chronic Lung Disease	*Chronic Lung Disease, if any in the Patient*
Balanced Diet	*If the patient is having a Balanced Diet or not*
Obesity	*Obesity Level in the Patient in the Patient*
Smoking	*Smoking habits of the patient*
Passive Smoker	*Passive Smoker if the patient has quit smoking*
Chest Pain	*Chest Pain, if any*
Coughing of Blood	*Coughing of Blood, if any*
Fatigue	*Fatigue level*
Weight Loss	*Sudden Weight Loss in the recent period*
Shortness of Breath	*Shortness of Breath, if any*
Wheezing	*Wheezing, if any*
Swallowing Difficulty	*Swallowing Difficulty, if any*
Clubbing of Finger Nails	*Clubbing of fingernails, if any*
Frequent Cold	*Frequency of occurrence of Cold*
Dry Cough	*Dry Cough or not*
Snoring	*If the patient snores*
Diagnosis	*The diagnosis of the Patient*

Table: 1 Basic Features of Lung Cancer

Features	Meaning
Sample Patient Number	*Id of the Patient*
Clump Thickness	*Clump Thickness of the cells*
Uniformity of Cell Size	*Uniformity of Cell Size*
Uniformity of Cell Shape	*Uniformity of Cell Shape*
Marginal Adhesion	*Marginal Adhesion of cell clumping*
Single Epithelial Cell Size Bare Nuclei	*Single Epithelial Cell Size in Bare Nuclei*
Bare Nuclei	*Bare Nuclei size*
Bland Chromatin	*Bland Chromatin*
Normal Nucleoli	*Normal Nucleoli or not*
Mitoses	*Mitoses occurrence*
Level	*Indicates Diagnosis of the Patient*

Table: 2 Basic Features of Breast Cancer

Features	**Meaning**
• patient_code • Diagnosis • behavior_sexualRisk • behavior_eating • behavior_personalHygine • intention_aggregation • intention_commitment • attitude_consistency • attitude_spontaneity • norm_significantPerson • norm_fulfillment • perception_vulnerability • perception_severity • motivation_strength • motivation_willingness • socialSupport_emotionality • socialSupport_appreciation • socialSupport_instrumental • empowerment_knowledge • empowerment_abilities • empowerment_desires	• *Patient ID details* • *Diagnosis of the patient* • *Sexual Risk of the patient* • *Eating Problems of the Patient* • *Personal Hygiene of the Patient* • *Aggregation* • *Commitment* • *Consistency of the Pain* • *Spontaneity of the Pain* • *Significant Person of the Patient* • *Fullness of Cysts, if any* • *Vulnerability to Cancer* • *Severity of the Cysts* • *Physical strength* • *Willingness* • *Mood Swings* • *Appreciation* • *Instrumental checkup (IVF)* • *Knowledge of Cancer* • *Fertility* • *Arousal of the Patient*

Table: 3 Basic Features of Cervical Cancer

In data preprocessing, specifically for feature selection, we typically use different techniques to determine the most influential features. The following are some techniques used here:

1. **Principal Component Analysis (PCA):**
   - PCA is a dimensionality reduction technique that aims to capture the most important information in the dataset by projecting it onto a lower-dimensional subspace.
   - Principal components (PCs) are calculated as linear combinations of the original features. The variance explained by each PC indicates its importance.
   - The cumulative explained variance ratio of PCs can be used to determine the proportion of total variance retained by the selected features.
2. **L1 Regularization (Lasso):**
   - L1 regularization penalizes the absolute magnitude of the coefficients in a linear model.
   - The regularization term added to the loss function is:

$$Regularization\,Term = \lambda \sum |\theta_i|$$

   - Features with non-zero coefficients after regularization are considered influential.

The following lists the various features selected for different cancer datasets.

Cancer Dataset	Features Selected
Lung	• Age • Gender • Air Pollution • Alcohol use • Dust Allergy • Occupational Hazards • Genetic Risk • Chronic Lung Disease • Obesity • Smoking • Diagnosis
Breast	• Clump Thickness • Uniformity of Cell Size • Uniformity of Cell Shape • Marginal Adhesion • Single Epithelial Cell Size Bare Nuclei • Bare Nuclei • Bland Chromatin • Normal Nucleoli • Mitoses • Level
Cervical	• behavior_sexualRisk • behavior_eating • behavior_personalHygine • perception_vulnerability • perception_severity • motivation_strength

Table: 4 Features Selected from Various Datasets

Splitting Data

80% of the dataset is taken as a training dataset used to train the machine. Once it is trained, we check the accuracy of the outcomes using a testing dataset.

$$X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, Y, test\_size = 0.2)$$

Algorithms

The following are the different algorithms implemented:

1. **Logistic Regression:**

   Logistic Regression is a statistical method used for modelling binary or multi-class classification problems. Despite its name, it is a linear classification model rather than a regression. It predicts the probability that an instance belongs to a particular class using the logistic function, which maps any real-valued number into the range [0, 1].

$$p(X; b, w) = \frac{1}{1 + e^{-w.X+b}}$$

2. **Naïve Bayes:**

   Naïve Bayes is a probabilistic classification algorithm based on Bayes' Theorem with the assumption of independence between features. It is particularly effective for text classification tasks such as spam detection and document categorization. Despite its simplicity and the "naïve" assumption, Naïve Bayes often performs well in practice and is computationally efficient.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

3. **Decision Tree:**

   A Decision Tree is a non-parametric supervised learning algorithm used for both classification and regression tasks. It recursively splits the dataset into subsets based on the feature that provides the most information gain or Gini impurity reduction. The final result is a tree-like structure where each internal node represents a feature, each branch represents a decision, and each leaf node represents the outcome.

$$p(k) = \frac{1}{n}\sum I(y = k)$$

4. **Support Vector Machine (SVM):**

   Support Vector Machine is a powerful supervised learning algorithm for classification, regression, and outlier detection tasks. SVM aims to find the hyperplane that best separates the classes in the feature space with the maximum margin. In cases where the data is not linearly separable, SVM uses the kernel trick to map the input features into a higher-dimensional space where separation is possible.

$$Linear\colon\ K(w, b) = w^T x + b$$
$$Polynomial\colon\ K(w, x) = (\gamma w^T x + b)^N$$
$$Gaussian\ RBF\colon\ K(w, x) = \exp\,(-\gamma||x_i - x_j||^n$$
$$Sigmoid\colon\ K(x_i, x_j) =\ \tanh(\alpha x_i^T x_j + b)$$

Performance Comparison

We have compared different algorithms such as Logistic Regression, Decision Tree, Naïve Bayes and SVM to choose the most efficient algorithm.

The following comparison table shows the accuracies of different ML algorithms:

ML Algorithm	**Lung Accuracy (%)**	**Breast Accuracy (%)**	**Cervical Accuracy (%)**
Logistic Regression	66	52	70
Naïve Bayes	87	74	83
Decision Tree	90	89	95
SVM	99	97	98

Table: 5 Comparative Analysis of ML Algorithms

Selecting Final Algorithm

Outputs for different Datasets:

Lung	**Breast**	**Cervical**
low	0	1
high	1	2

Table: 6 Different Outputs for Datasets

All these different outputs for the same categories are defined using single variables. The outputs indicating a low level of cancer risk such as “Low”, “0”, and “1” are assigned the variable name “Low Risk”. Similarly, the variable “High Risk” is assigned to “High”, “1”, and “2” which indicate a high level of cancer risk.

Outputs	**Variable**
low, 0, 1	Low Risk
high, 1, 2	High Risk

Table: 7 Selected Output Variables

From this comparison, we have decided to use the Support Vector Machine (SVM), since it has the highest accuracy percentage.

Clinical Interpretation

SVM's results can be analysed to understand how effectively it can assist healthcare professionals in making treatment decisions for cancer patients. SVM's ability to handle complex, multi-dimensional data and provide accurate prognosis predictions may have a significant impact on patient outcomes and treatment planning in clinical practice.

SVM's flexibility, the ability to work with both linear and non-linear data, and its performance evaluation through various metrics make it a valuable tool in cancer prognosis prediction. The detailed analysis of SVM and its comparison with other algorithms provide a comprehensive assessment of its effectiveness in your study.

Conclusion

This study extensively explores machine learning algorithms for cancer prognosis prediction, specifically targeting lung, breast, and cervical cancers. Through rigorous methodology encompassing data collection, preprocessing, feature selection, and algorithm implementation, the research aims to enhance prognostic accuracy and clinical relevance.

The findings highlight the superiority of Support Vector Machines (SVM) in accurately predicting cancer prognosis across diverse datasets. SVM's effectiveness stems from its ability to handle complex data and provide precise predictions, potentially revolutionizing treatment decisions and patient outcomes in clinical settings.

Furthermore, the study emphasizes the importance of integrating clinical, molecular, and cytological data to improve prognostic accuracy. While the presented models exhibit promise, further validation on larger datasets is necessary to ensure their broader applicability.

In summary, this research underscores the transformative potential of advanced computational methodologies in cancer diagnostics. Despite limitations, such as data availability, the integration of AI-driven techniques with medical expertise holds promise for enhancing cancer detection and treatment, ultimately benefiting patient care and outcomes.

References

[1] KwetisheJoroDanjuma," Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients" Department of Computer Science, ModibboAdama University of Technology, Yola, Adamawa State, Nigeria

[2] Survey of Intelligent Methods for Brain Tumor Detection-IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 1, September 2014

[3] Zehra Karhan1, Taner Tunç2," Lung Cancer Detection and Classification with Classification Algorithms" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 22788727, Volume 18, Issue 6, Ver. III (Nov.-Dec. 2016), PP 71-77.

[4] Ada, RajneetKaur," A Study of Detection of Lung Cancer Using Data Mining Classification Techniques" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013

[5] R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, "Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis," IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), vol. 4, no. 3, pp. 485-495, 2007

[6] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," Computer methods and programs in biomedicine, vol. 153, pp. 1-9, 2018.