Veri Bilim Derg, 6(2), 24-40, 2023

VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Gelişmiş Orofarengeal Kanser Prognozu: Yenilikçi Bir Toplululuk Makine Öğrenimi Yaklaşımı

Pınar KARADAYI ATAŞ[1]*

[1]*İstanbul Arel University, Faculty of Engineering and Architecture, Computer Engineering, İstanbul*

Makale Bilgisi

Başvuru:
05/12/2023
Kabul:
16/12/2023

Özet

Orofaringeal kanserli hastalarda hayatta kalma oranlarını tahmin etmek için makine öğrenimi algoritmalarının kullanılması bu çalışmanın ana odak noktasıdır. Kanser prognozunun doğasında olan karmaşıklık ve değişkenlik göz önüne alındığında, geleneksel öngörücü modeller genellikle doğruluk ve güvenilirlik açısından yetersiz kalmaktadır. Gaussian Naive Bayes, Rastgele Orman, Gradyan Arttırma, Doğrusal Destek Vektör Makinesi, Lojistik Regresyon ve K-En Yakın Komşular dahil olmak üzere, bu sorunların üstesinden gelmek için her biri veri analizinde kendi avantajları olan çeşitli makine öğrenimi yöntemlerini kullandık. Bu ayrı algoritmaları birleştiren bir topluluk modelinin geliştirilmesi stratejimizin anahtarıydı. Bu modelin genel tahmin gücü, tüm tekniklerin birleşik avantajlarından yararlanılarak artırılmıştır. Karşılaştırmalı analizimizin sonuçları, bireysel algoritmaların performansı farklılık gösterse de önerilen topluluk modelinin hepsinden daha iyi performans gösterdiğini, daha yüksek doğruluk (0.92), f1 puanı (0.92), kesinlik (0.93) ve duyarlılık (0.90) elde ettiğini gösterdi. Çalışmanın bulguları, özellikle tıbbi teşhis için kanser prognozunun karmaşık alanında toplu makine öğrenimi modellerinin potansiyelini vurgulamaktadır. Topluluk modeli, birden fazla algoritmayı verimli bir şekilde birleştirerek orofaringeal kanserli hastalarda hayatta kalma sonuçlarını tahmin etmek için daha kapsamlı bir araç sunar. Bu yöntem sadece öngörü doğruluğunu arttırmakla kalmıyor, aynı zamanda hastalığın dinamiklerinin daha derinlemesine anlaşılmasını sağlayarak daha bireysel ve başarılı tedavi planlarının kapısını aralıyor.

Anahtar Kelimeler: *Veri Bilimi, Orofaringeal Kanser, Tıbbi Teşhis, Makine Öğrenimi.*

Advancing Oropharyngeal Cancer Prognosis: A Novel Ensemble Machine Learning Approach

Abstract

The use of machine learning algorithms to forecast survival rates in patients with oropharyngeal cancer is the main focus of this study. Given the complexity and variability inherent in cancer prognosis, traditional predictive models often fall short in accuracy and reliability. We used a variety of machine learning methods, each with its own advantages in data analysis, to tackle these problems, including Gaussian Naive Bayes, Random Forest, Gradient Boosting, Linear Support Vector Machine, Logistic Regression, and K-Nearest Neighbors. The development of an ensemble model that combined these separate algorithms was the key to our strategy. The overall predictive power of this model is increased by utilizing the combined advantages of all the techniques. The results of our comparative analysis indicated that although the performance of the individual algorithms varied, the suggested

---

* Iletişim e-posta: pinaratas@arel.edu.tr

ensemble model performed better than all of them, obtaining higher accuracy (0.92), f1-score (0.92), precision (0.93), and recall (0.90). The study's findings highlight the potential of ensemble machine learning models in the complex field of cancer prognosis in particular, for medical diagnostics. The ensemble model offers a more comprehensive tool for predicting survival outcomes in patients with oropharyngeal cancer by efficiently combining multiple algorithms. This method not only increases the predictive accuracy but also provides a deeper comprehension of the dynamics of the disease, opening the door to more individualized and successful treatment plans.

Keywords: *Data Science, Oropharyngeal Cancer, Medical Diagnostics, Machine Learning.*

1 Introduction

The use of machine learning (ML) in healthcare, especially in oncology, has emerged as a ray of hope for improving prognostic evaluations and diagnostic accuracy in the fast-moving field of precision medicine [1]. The prediction of survival outcomes for patients with oropharyngeal cancer is the main focus of this paper's groundbreaking investigation into the field of machine learning [2]. The oropharynx, or the area of the throat directly behind the mouth that includes the tonsils and base of the tongue, is primarily affected by oropharyngeal cancer, a type of cancer that is well-known for its difficult treatment dynamics and late diagnosis. When it comes to effective treatment, timely diagnosis is crucial because symptoms like persistent sore throat, unexplained weight loss, and hoarseness can delay treatment[3], [4] .

An unsettling increase in the incidence of oropharyngeal squamous cell carcinoma (OPSCC), a dangerous rival in the field of head and neck cancers, has been observed, especially in developed nations [5]. Making up a significant percentage of the 100,000 new cases that are diagnosed each year, this type of cancer has particular difficulties and complications in managing and treating it [6]. The causes of OPSCC have changed over time, with the human papillomavirus (HPV) emerging as a major modifiable risk factor in addition to more conventional causes like smoking and heavy alcohol consumption [7], [8]. Compared to traditional upper respiratory airway squamous cell carcinomas, HPV-associated OPSCC notably affects a younger population, indicating a paradigm shift in the field of head and neck carcinomas.

Surgery, radiation therapy, and chemotherapy are just a few of the aggressive multimodal treatments used to treat OPSCC, particularly in cases where the cancer is linked to HPV [6] [9]. Although necessary, this aggressive treatment can have a significant negative impact on patients' quality of life, as the aftereffects of such interventions may negatively impact their survival for decades. Thus, to minimize these side effects and guarantee successful cancer management, individualized treatment regimens that are based on each patient's unique risk profile are desperately needed.

Patient survival can be increased by more effectively customizing treatment and controlling the progression of the disease with the help of a trustworthy, validated cancer prognosis [10]. These days, medical images (such as CT, MR, or PET images) can have massive quantitative image features mined from them. The most valuable image features can then be filtered out using statistical or machine learning techniques for purposes such as disease characterization, prognosis prediction, tumor grading and staging, efficacy assessment, and clinical information parsing.

Different machine learning (ML) techniques have been applied extensively in cancer prognosis in recent years to increase prediction accuracy because they are more accurate and efficient than conventional statistical analysis techniques [11]. According to [12], the accuracy of cancer prediction results has increased by 15% to 20% with the use of ML technology. Among them, Tseng et al. [13] predicted the 5-year survival of patients with oral cancer using logistic regression, and the accuracy was 0.68. Chu et al. [14]predicted treatment outcomes for oral cancer using decision tree, linear regression, and k-nearest neighbor models; the area-under-ROC-curve (AUC) preferably reached 0.67, 0.7, and 0.71, respectively. AUC for the random forest model used by Patel et al. [15] to forecast the overall survival of OPC patients was 0.80. Deep learning (DL) models are gaining popularity in addition to traditional machine learning models. Cheng et al. [16] created a deep learning model with an AUC of 0.801 for predicting overall survival in OPC patients. These studies do have certain issues, though. Due to their inherent simplicity, simple machine learning models like linear regression and logical regression decision trees can produce easily comprehensible prediction results; however, their prediction accuracy is notably lacking [17]. More people are using sophisticated, high-precision machine learning models, like DL models and tree-based ensemble models.

Radiomics is the term for this procedure [18], [19]. Prognostic relationships between images and medical outcomes may become apparent as a result of this deeper mining of medical image feature data [20]. To create models that could potentially increase diagnostic, prognostic, and predictive accuracy, radiomic data are frequently combined with other patient data [21]. As a result, radiomics has excellent application potential in the OPC prognosis. However, because of the growing complexity of cancers, treatment plans, and patient demographic samples, it is still difficult to accurately predict survival in OPC cancer patients[22], [23].

Our research sets out on a thorough investigation, assessing a variety of machine learning algorithms, each with its own distinct approach and capacity for prediction Gaussian Naive Bayes, Random Forest, Gradient Boosting, Linear-SVM, Logistic Regression, and K-Nearest Neighbors (KNN) are among the algorithms that this study examines in detail. These algorithms were carefully selected because of their track record of success in a variety of medical applications, especially cancer prognosis. The aim of this study is to evaluate and contrast their abilities in forecasting oropharyngeal cancer patients' survival rates, a critical task that affects clinical decision-making.

In this paper, we investigate further the integration of these algorithms via advanced ensemble classification techniques. This strategy aims to counteract any shortcomings and increase overall prediction accuracy by combining the benefits of various algorithms. The application of machine learning algorithms to patient outcome prediction in the context of oropharyngeal cancer, where such predictive models are uncommon, is a novel approach in medical research. Furthermore, this study presents novel visual analytics instruments aimed at improving the comprehensibility of intricate algorithmic results. These resources are essential for giving researchers and clinicians a clear understanding of the predictive behaviors of various algorithms. In addition to improving comprehension of the data, the visualization of algorithmic predictions and performance metrics also helps in the critical assessment and choice of the best algorithm for clinical use. This work is important because it has the potential to completely change how oropharyngeal cancer patients are diagnosed and treated. Through the identification of the optimal machine learning algorithm, or combination thereof, this study seeks to make a substantial contribution to the oncology field. The study's predictive models may play a significant role in helping to guide treatment choices, customize patient care, and eventually raise the quality of life and survival rates for oropharyngeal cancer patients.

2 Materail and Methods

2.1 Data Set

The dataset used in this study came from the Cancer Imaging Archive “https://www.cancerimagingarchive.net/”, a reputable source of clinical datasets. This dataset consists of 607 individual records, each of which is defined by 34 different attributes. It offers a comprehensive picture of patients with oropharyngeal cancer and was painstakingly assembled by licensed experts. The characteristics present in the dataset play a crucial role in representing the survival results of these patients. Making use of this extensive dataset, we have compared several machine learning algorithms. The main goal was to assess how well these algorithms performed in correctly predicting survival outcomes from the given attributes, which would provide important information for the management and treatment of oropharyngeal cancer.

2.2 Data Preprocessing

In order to prepare the dataset for machine learning models, data preprocessing was an essential step in this study. To prepare the dataset for analysis, the first step of the process involved locating and eliminating features that were unnecessary. There were 606 entries in the original dataset, each with 34 attributes. After removing features and missing data, the refined dataset was reduced to 606 entries with 22 attributes. This step was essential in guaranteeing the data's relevance and integrity, which improved the machine learning models' accuracy and dependability when used in the analysis that followed.

The mentioned The distribution of relationships between the 'Status' label column and three different numerical attributes is shown in the study's Figures 1 and 2. These graphs provide light on the patterns and trends found in the dataset by clearly illustrating the relationship between each of these particular characteristics and the patients' "Status." These kinds of visualizations are essential for comprehending the dynamics that lie beneath the data and for providing direction when machine learning algorithms are applied later on to forecast patient outcomes.

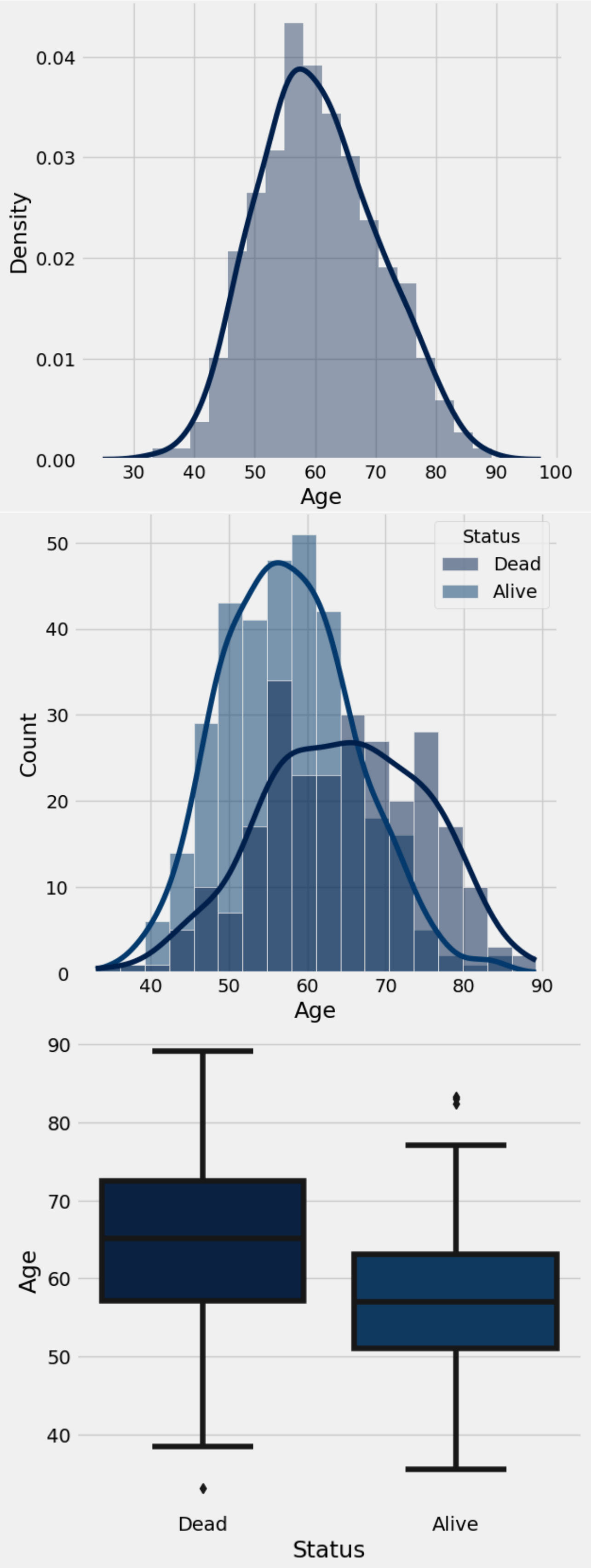


Figure 1. Relation between age and status.

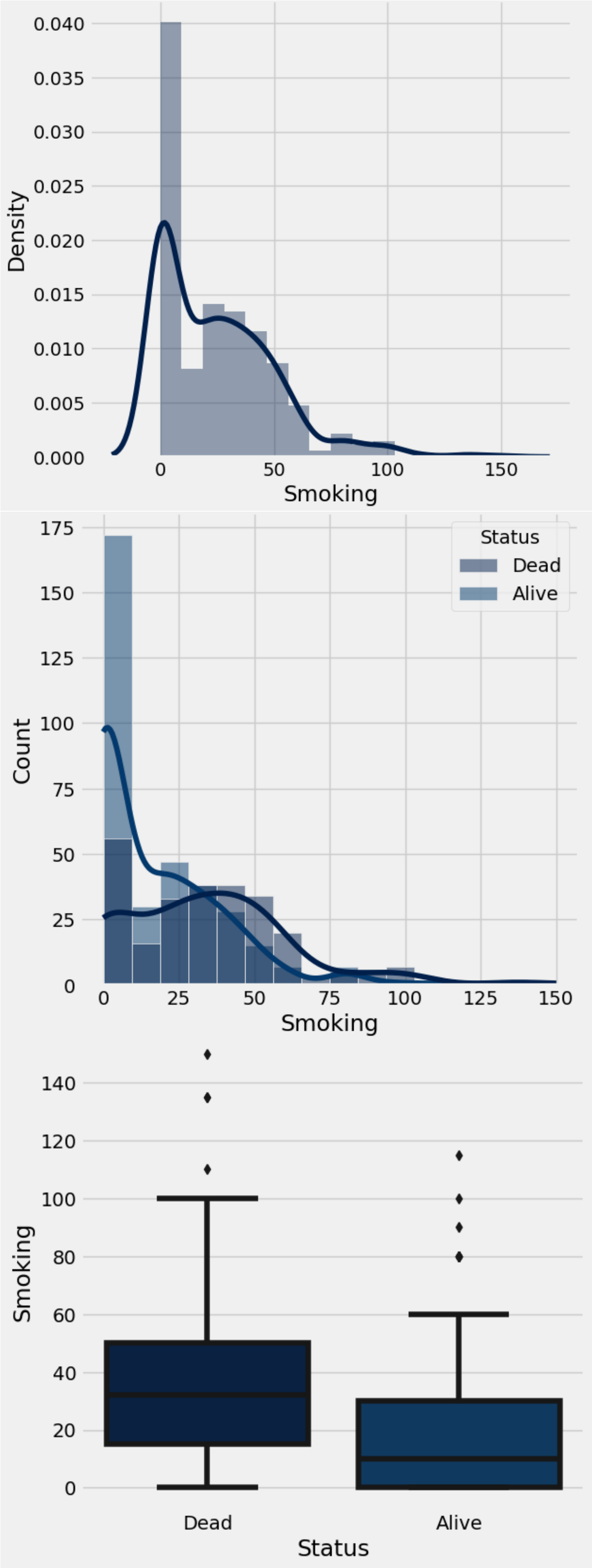


Figure 2. Relation between smoking and status.

The distribution of relationships between the 'Status' label column and three different categorical attributes is the main focus of Figures 3 and 4.. The relationships and interactions between each category attribute and the patients' "Status" are crucially shown by these figures. The figures offer important insights into how various categorical factors may affect patient outcomes by visualizing these relationships. This knowledge is crucial for the efficient use of machine learning algorithms in the research since it facilitates the identification of important predictors and the customization of the algorithms to better capture the subtleties of the dataset.

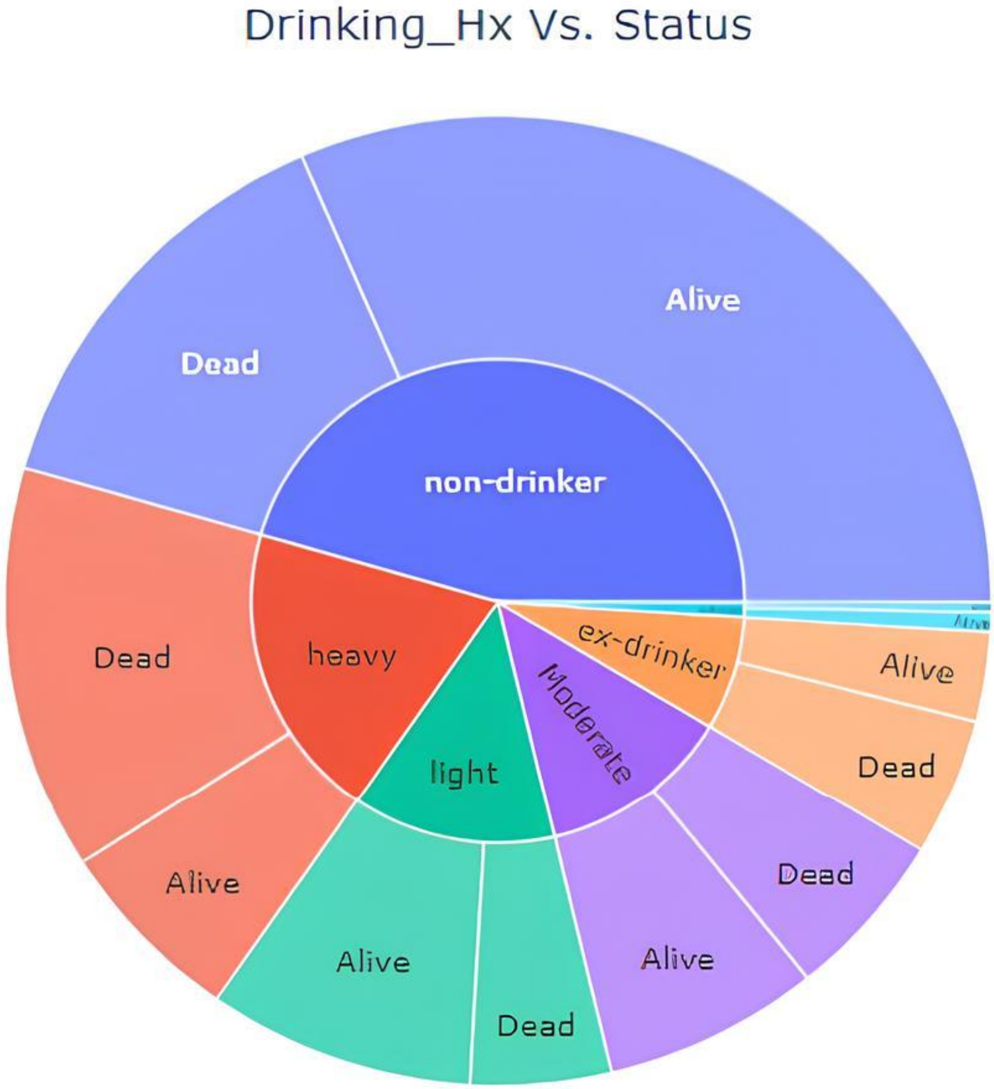


Figure 3. The relationship between 'Status' and 'Drinking History' (Drinking_Hx).

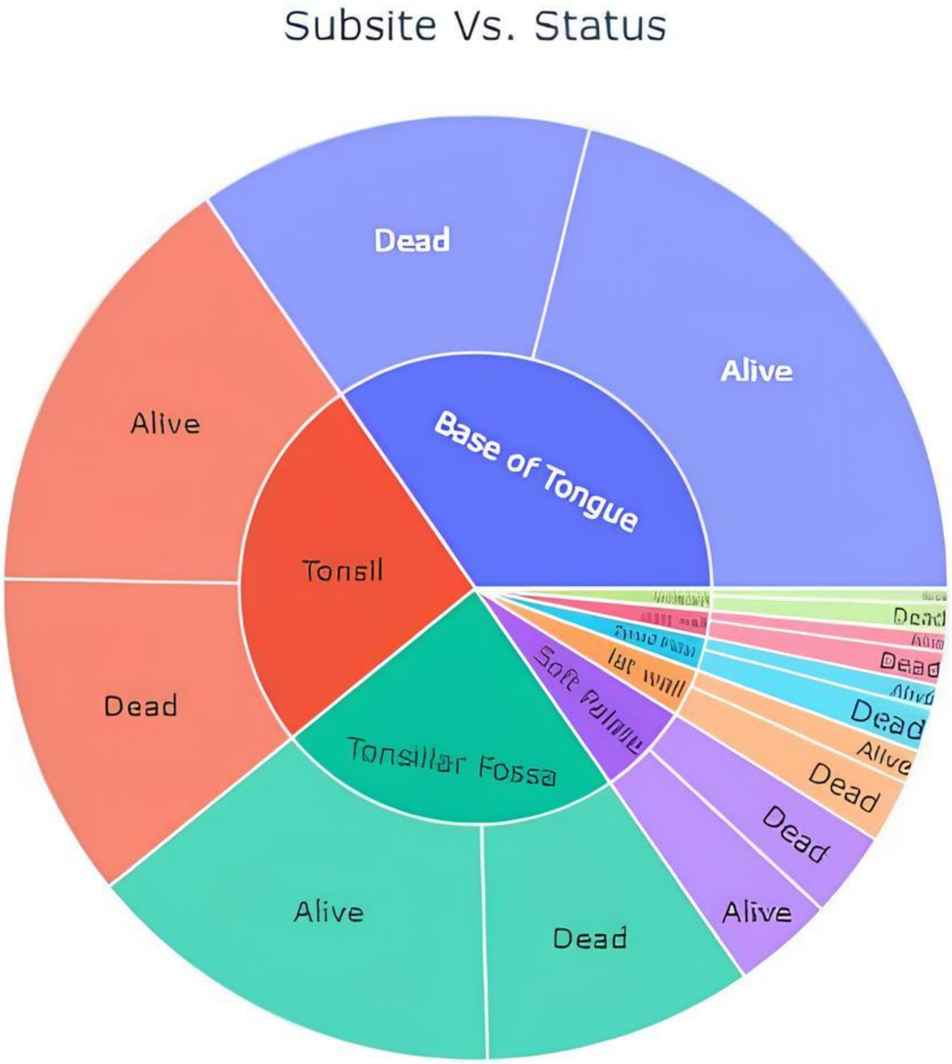
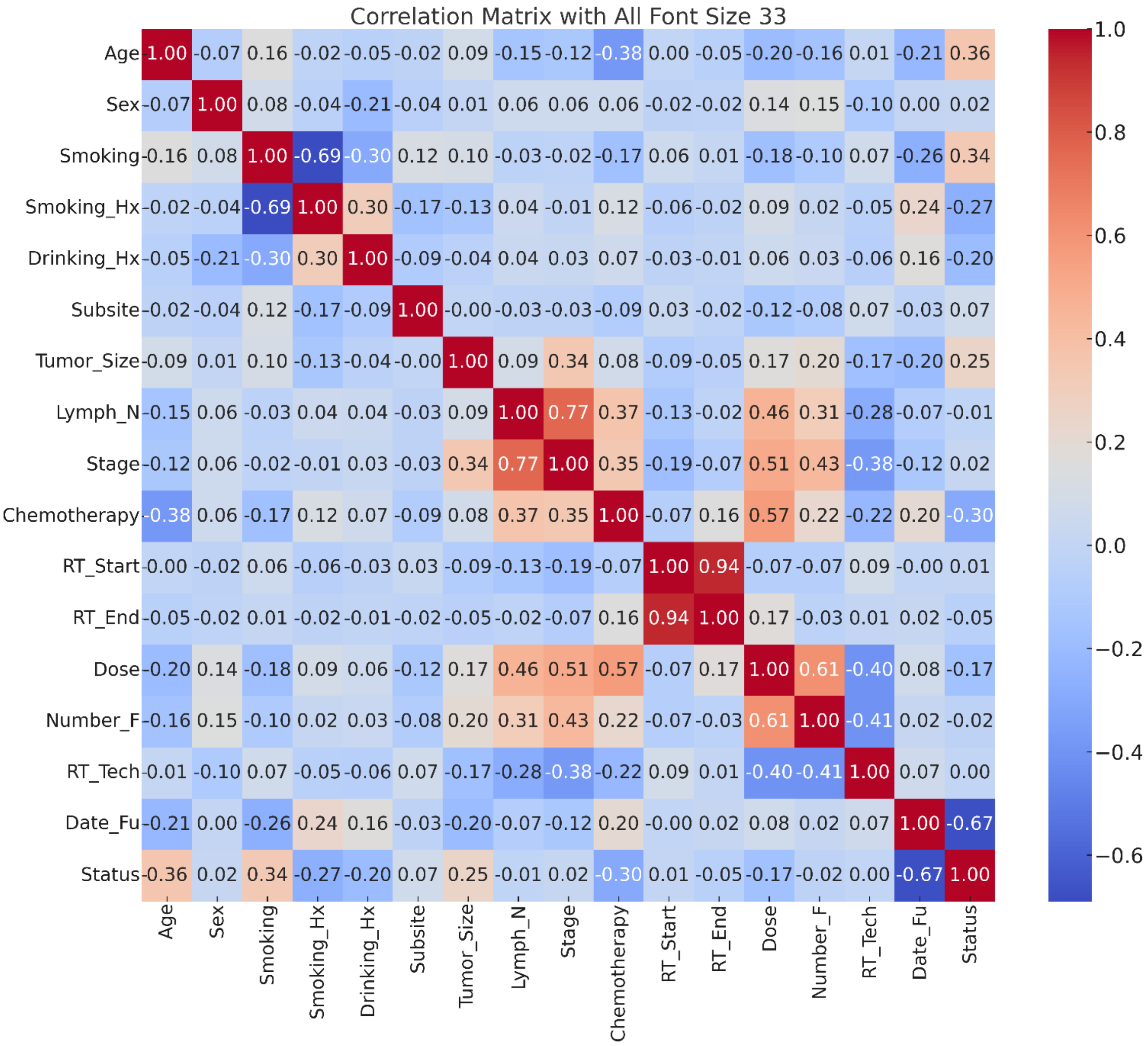


Figure4. Relation between Status and Subsite.

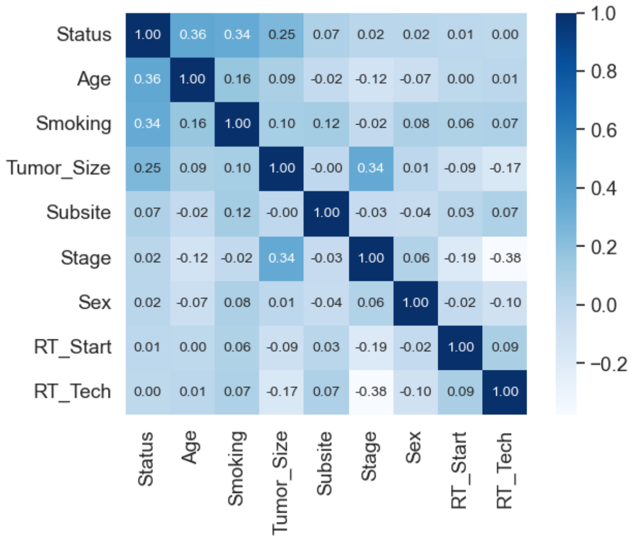
Understanding the relationships between variables in a multivariate dataset is made easier with the help of the correlation matrix, which is an essential tool in statistical analyses and data exploration procedures. A strong relationship with the label is indicated by values in the correlation matrix that are close to 1, whereas a weak or insignificant relationship is indicated by values that are close to 0.



	Age	Sex	Smoking	Smoking_Hx	Drinking_Hx	Subsite	Tumor_Size	Lymph_N	Stage	Chemotherapy	RT_Start	RT_End	Dose	Number_F	RT_Tech	Date_Fu	Status
Age	1.00	-0.07	0.16	-0.02	-0.05	-0.02	0.09	-0.15	-0.12	-0.38	0.00	-0.05	-0.20	-0.16	0.01	-0.21	0.36
Sex	-0.07	1.00	0.08	-0.04	-0.21	-0.04	0.01	0.06	0.06	0.06	-0.02	-0.02	0.14	0.15	-0.10	0.00	0.02
Smoking	0.16	0.08	1.00	-0.69	-0.30	0.12	0.10	-0.03	-0.02	-0.17	0.06	0.01	-0.18	-0.10	0.07	-0.26	0.34
Smoking_Hx	-0.02	-0.04	-0.69	1.00	0.30	-0.17	-0.13	0.04	-0.01	0.12	-0.06	-0.02	0.09	0.02	-0.05	0.24	-0.27
Drinking_Hx	-0.05	-0.21	-0.30	0.30	1.00	-0.09	-0.04	0.04	0.03	0.07	-0.03	-0.01	0.06	0.03	-0.06	0.16	-0.20
Subsite	-0.02	-0.04	0.12	-0.17	-0.09	1.00	-0.00	-0.03	-0.03	-0.09	0.03	-0.02	-0.12	-0.08	0.07	-0.03	0.07
Tumor_Size	0.09	0.01	0.10	-0.13	-0.04	-0.00	1.00	0.09	0.34	0.08	-0.09	-0.05	0.17	0.20	-0.17	-0.20	0.25
Lymph_N	-0.15	0.06	-0.03	0.04	0.04	-0.03	0.09	1.00	0.77	0.37	-0.13	-0.02	0.46	0.31	-0.28	-0.07	-0.01
Stage	-0.12	0.06	-0.02	-0.01	0.03	-0.03	0.34	0.77	1.00	0.35	-0.19	-0.07	0.51	0.43	-0.38	-0.12	0.02
Chemotherapy	-0.38	0.06	-0.17	0.12	0.07	-0.09	0.08	0.37	0.35	1.00	-0.07	0.16	0.57	0.22	-0.22	0.20	-0.30
RT_Start	0.00	-0.02	0.06	-0.06	-0.03	0.03	-0.09	-0.13	-0.19	-0.07	1.00	0.94	-0.07	-0.07	0.09	-0.00	0.01
RT_End	-0.05	-0.02	0.01	-0.02	-0.01	-0.02	-0.05	-0.02	-0.07	0.16	0.94	1.00	0.17	-0.03	0.01	0.02	-0.05
Dose	-0.20	0.14	-0.18	0.09	0.06	-0.12	0.17	0.46	0.51	0.57	-0.07	0.17	1.00	0.61	-0.40	0.08	-0.17
Number_F	-0.16	0.15	-0.10	0.02	0.03	-0.08	0.20	0.31	0.43	0.22	-0.07	-0.03	0.61	1.00	-0.41	0.02	-0.02
RT_Tech	0.01	-0.10	0.07	-0.05	-0.06	0.07	-0.17	-0.28	-0.38	-0.22	0.09	0.01	-0.40	-0.41	1.00	0.07	0.00
Date_Fu	-0.21	0.00	-0.26	0.24	0.16	-0.03	-0.20	-0.07	-0.12	0.20	-0.00	0.02	0.08	0.02	0.07	1.00	-0.67
Status	0.36	0.02	0.34	-0.27	-0.20	0.07	0.25	-0.01	0.02	-0.30	0.01	-0.05	-0.17	-0.02	0.00	-0.67	1.00

Figure 5. Correlation Matrix.

In Figure 5 effectively visualizes the strength of these relationships with a heatmap of the top 9 values showing the highest correlation with our label. In order to further explore this analysis, Figure 6 shows the bivariate relationships by displaying these attributes in the form of a scatter plot. This method aids in both determining the important variables that significantly affect the label and comprehending the nature of their interactions.



	Status	Age	Smoking	Tumor_Size	Subsite	Stage	Sex	RT_Start	RT_Tech
Status	1.00	0.36	0.34	0.25	0.07	0.02	0.02	0.01	0.00
Age	0.36	1.00	0.16	0.09	-0.02	-0.12	-0.07	0.00	0.01
Smoking	0.34	0.16	1.00	0.10	0.12	-0.02	0.08	0.06	0.07
Tumor_Size	0.25	0.09	0.10	1.00	-0.00	0.34	0.01	-0.09	-0.17
Subsite	0.07	-0.02	0.12	-0.00	1.00	-0.03	-0.04	0.03	0.07
Stage	0.02	-0.12	-0.02	0.34	-0.03	1.00	0.06	-0.19	-0.38
Sex	0.02	-0.07	0.08	0.01	-0.04	0.06	1.00	-0.02	-0.10
RT_Start	0.01	0.00	0.06	-0.09	0.03	-0.19	-0.02	1.00	0.09
RT_Tech	0.00	0.01	0.07	-0.17	0.07	-0.38	-0.10	0.09	1.00

Figure 6. Heat Map of the Top 9 Features Closest to the Correlation Matrix.

2.3 Methods

This work sets out on an extensive analytical journey to predict the survival outcomes of patients with oropharyngeal cancer using a wide range of machine learning algorithms. A variety of machine learning algorithms have been used in this study, carefully chosen for their individual merits and applicability to the challenging task of predicting survival outcomes in patients with oropharyngeal cancer. Because of its effectiveness with big datasets and resilience in prediction tasks with little data, Gaussian Naive Bayes is widely used. This makes it perfect for managing the complex patterns found in medical datasets. Because of their great accuracy and versatility in handling a wide range of input variables, Random Forest and Gradient Boosting are included in our extensive dataset. Additionally, by lowering overfitting, these ensemble techniques improve the model's capacity for generalization. The selection of Support Vector Machine (SVM) stems from its efficacy in high-dimensional spaces and its potential to represent non-linear relationships, which are frequently found in medical data. Because it is easy to understand and effective in binary classification, logistic regression is chosen to provide precise probabilistic insights that are crucial for making medical decisions. In conclusion, K-Nearest Neighbors (KNN) is incorporated due to its ease of use and efficiency in classification assignments, providing significant insight in medical situations where comprehending nearby data trends is essential for making patient-specific forecasts. When combined, these algorithms offer a strong, comprehensive method for precisely predicting patient outcomes, which is crucial for individualized treatment planning.

The simplicity and efficacy of Gaussian Naive Bayes [24], especially when dealing with categorical data, led to its selection. The underlying premise of this algorithm is the predictors' independence. Because of its effectiveness in handling big datasets and its resilience in generating predictions with little data, it is especially helpful in this study. Medical datasets frequently contain complex patterns that can be analyzed using Gaussian Naive Bayes, which is well-known for its application in classification tasks and its capacity to handle high-dimensional data.

An ensemble learning technique called Random Forest works by building a large number of decision trees during training [25]. It outputs the individual trees' mean prediction (regression) or the mode of the classes (classification). Because it can handle multiple input variables without deleting any of them—a critical feature in complex datasets such as those found in the medical domain. Random Forest is renowned for its high accuracy and efficiency when processing large datasets. Random Forest is renowned for its high accuracy, efficiency when processing large datasets, and capacity to handle multiple input variables without deleting any of them—a critical feature in complex datasets such as those found in the medical domain.

Gradient Boosting is a potent machine learning method that builds a model step-by-step for regression and classification problems [26]. To arrive at the final prediction, it builds new models that predict the residuals or errors of earlier models, which are then added together. It was selected because of its versatility in optimizing different loss functions and its capacity to handle a wide range of data types, making it an excellent fit for intricate survival analyses.

A strong and adaptable supervised machine learning algorithm, support vector machine (SVM) works especially well in high-dimensional spaces [27]. Its kernel trick, which raises linearly inseparable data to a higher dimension where it becomes separable, is its most well-known feature. SVM is selected because it works well with complicated medical datasets, where modeling non-linear relationships is essential.

In its most basic form, logistic regression is a statistical model that models a binary dependent variable using a logistic function [28]. It is frequently utilized for binary classification tasks, such as disease/no disease, in the context of medical data. It is selected due to its simple interpretation, simplicity of application, and the unambiguous probabilistic comprehension of the class labels it offers, all of which are critical for medical decision-making.

A straightforward instance-based learning algorithm is K-Nearest Neighbors (KNN). KNN produces a class membership as its output [29]. An object is classified by a majority vote of its neighbors, and it is placed in the class that has the greatest commonality among its k closest neighbors. KNN is selected due to its ease of use, efficacy, and intuitiveness; the latter being especially helpful in a clinical context where knowledge of local data patterns is essential for making patient-specific predic

2.4 Proposed Method

We used a novel approach in the study's proposed methodology to improve the prediction accuracy for the survival of oropharyngeal cancer. Initially, each row of the dataset was subjected to the Support Vector Machine (SVM) algorithm in order to determine the prediction probability for each individual entry. Subsequently, a threshold value was established by calculating the average of these probabilities.

This threshold was then used to split the dataset into two equal portions, one containing values higher than the threshold and the other containing values lower than the threshold. Stratification was a calculated move to customize the algorithms' application to distinct dataset segments. The idea behind this was to account for the possibility that distinct subsets would display distinct patterns that would be better represented by particular models.

A variety of algorithms, such as Gaussian Naive Bayes, Random Forest, Gradient Boosting, SVM, and K-Nearest Neighbors (KNN),and Logistic Regression were used for each segmented portion of the dataset. Ten models in all, five from each dataset segment, were produced as a result of this process. The implementation of an ensemble approach, supported by a voting mechanism, significantly improved the overall precision and dependability of our survival estimates for patients diagnosed with oropharyngeal cancer.

The creation of an ensemble model represented the method's apex. By combining the individual models' strengths and making up for any shortcomings, this ensemble increased the predictive power of each model. Our methodology concluded with the application of a voting algorithm to this ensemble. The voting algorithm eventually produced a thorough and reliable prediction for every case in the dataset by combining the predictions from each of the ten models. The algorithm exit in bellow.

Algorithm: Predictive Modeling for Oropharyngeal Cancer Survival

```
1: Initialize the Support Vector Machine (SVM) model.
2: For each row in the dataset:
    a. Apply the SVM model to obtain a prediction probability.
3: Calculate the average of all prediction probabilities from SVM.
4: Set this average as the threshold value.
5: Split the dataset into two parts:
    a. Part A: Rows with prediction probabilities greater than the threshold.
    b. Part B: Rows with prediction probabilities less than the threshold.
6: For each part (A and B):
    a. Apply the following algorithms:
        i. Gaussian Naive Bayes
        ii. Random Forest
        iii. Gradient Boosting
        iv. Support Vector Machine (SVM)
        v. K-Nearest Neighbors (KNN)
        vi. Logistic Regression
    b. This results in 6 models for each part, totaling 12 models.
7: Create an ensemble model comprising the 12 models from Parts A and B.
8: Apply a voting algorithm to the ensemble:
    a. Aggregate predictions from all 12 models in the ensemble.
```

b. Determine the final prediction based on the majority vote.

9: Output the final prediction for each case in the dataset.

3 Results

A detailed analysis is conducted to determine how well different machine learning algorithms predict the prognosis of patients with oropharyngeal cancer. A number of algorithms are examined, including Gaussian Naive Bayes, Random Forest, Gradient Boosting, Support Vector Machine (SVM), Logistic Regression and K-Nearest Neighbors (KNN). The study shows how effective these algorithms are and combines their benefits with hybrid classification techniques. This method integrates the Base Ensemble, Rule Ensemble, and Spline Rule Ensemble algorithms to produce a classifier that is more capable and adaptable.

Matplotlib and Plotly libraries were used to visually analyze the relationships between the data. These analyses contribute to our understanding of the variables and factors that affect oropharyngeal cancer patients' chances of survival. Visual analyses provided by Graphical User Interfaces (GUIs) facilitate data interpretation and result tracking in a more easily accessible and understandable manner. Graphics such as confusion matrices and ROC AUC curves are used to display performance metrics like specificity, accuracy, false positive rates, and false negative rates. These visual aids are essential for assessing how well the algorithms work and for streamlining decision-making procedures. Furthermore, one can monitor the algorithms' performance in relation to training data through learning curve analyses displayed on the graphical user interface. With the help of this feature, problems like overfitting or underfitting can be found and fixed, allowing the models to be more broadly applicable. The precision-recall curve helps users choose the right threshold for optimal performance by displaying the algorithms' sensitivity and recall values at various classification thresholds. The classification report makes it easier to compare algorithms and highlights the output of the most successful algorithm by giving a thorough summary of performance metrics like precision, sensitivity, F1 score, and support.

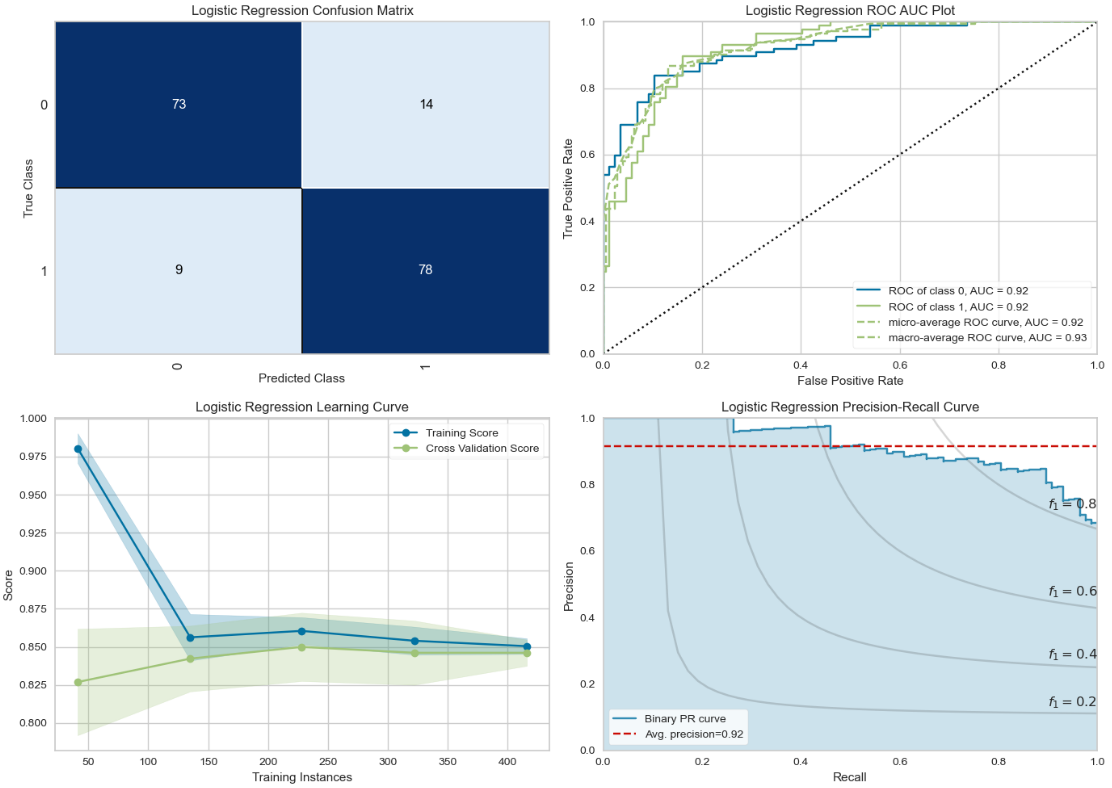


Figure 7. Logistic Regression Algorithm Results.

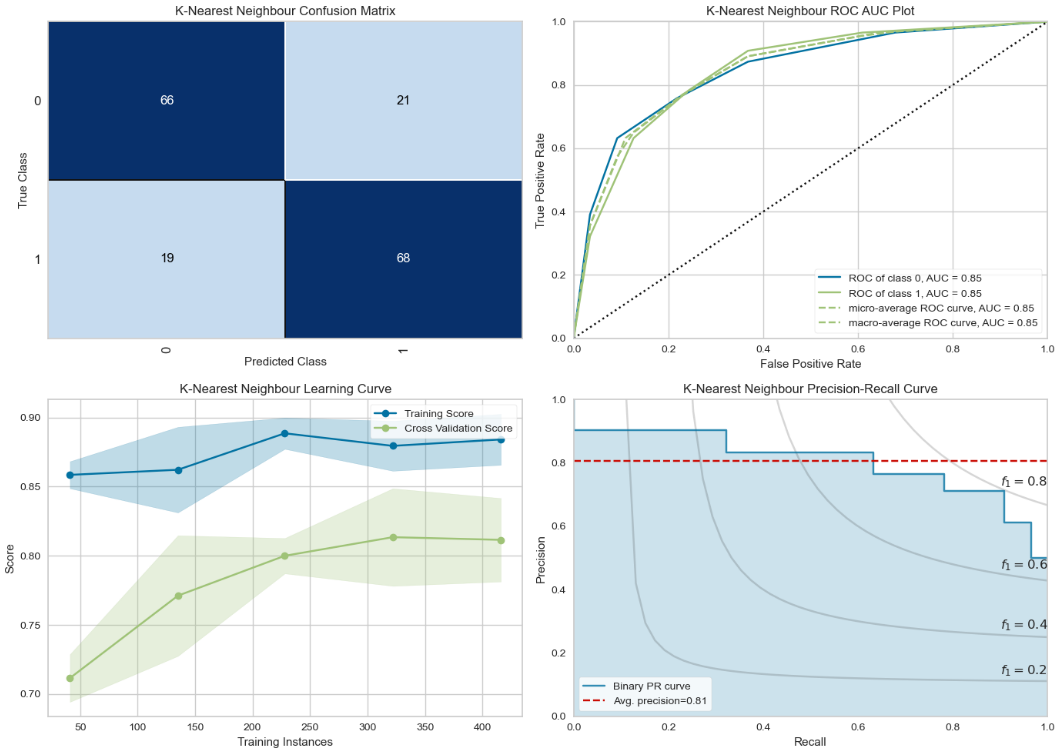


Figure 8. K-Nearest Neighbors (KNN)Algorithm Results.

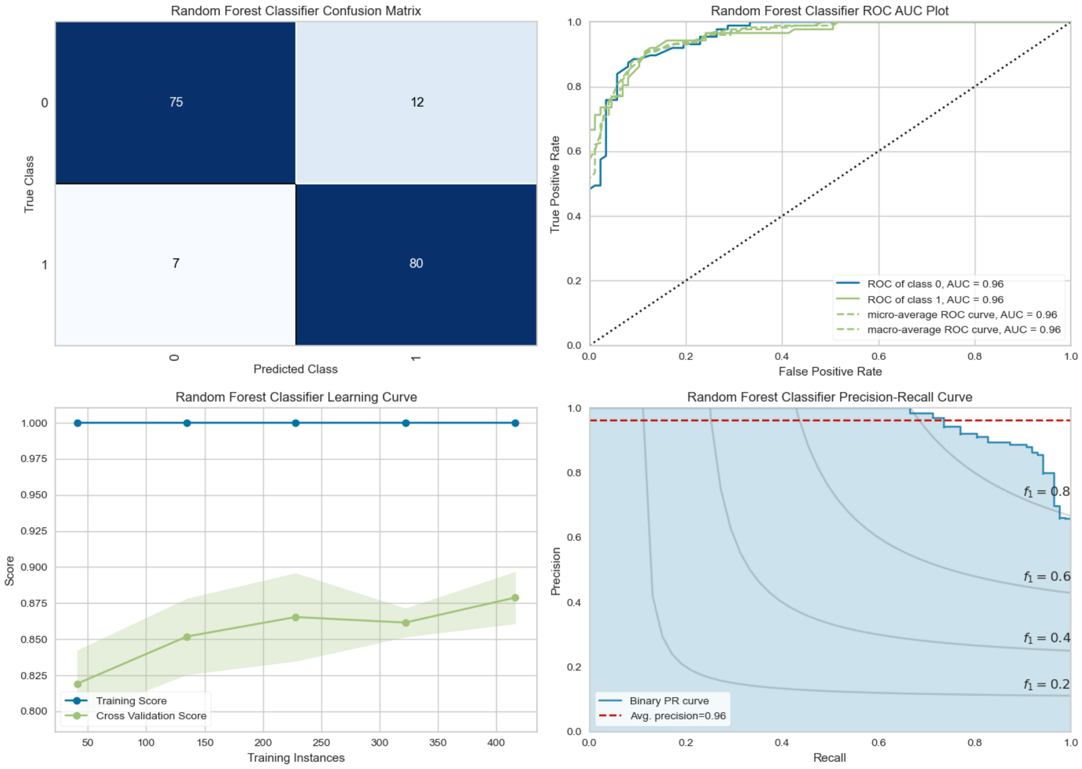


Figure 9. Random Forest Algorithm Results.

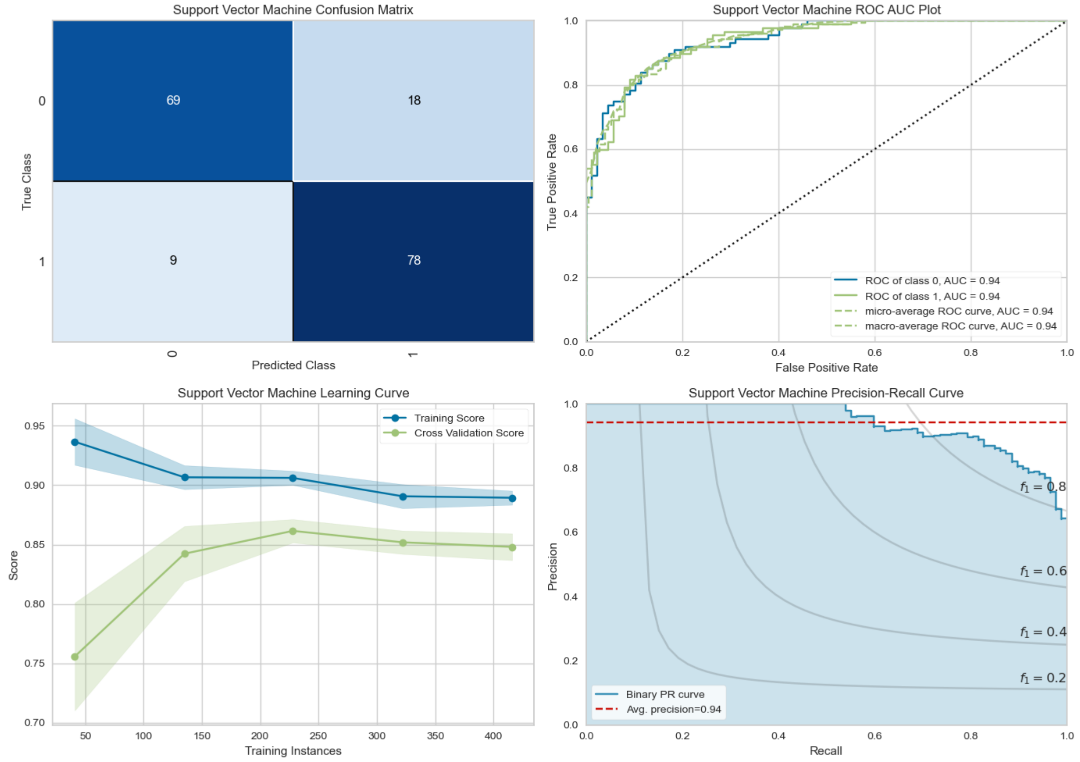


Figure 6. SVM Algorithm Results.

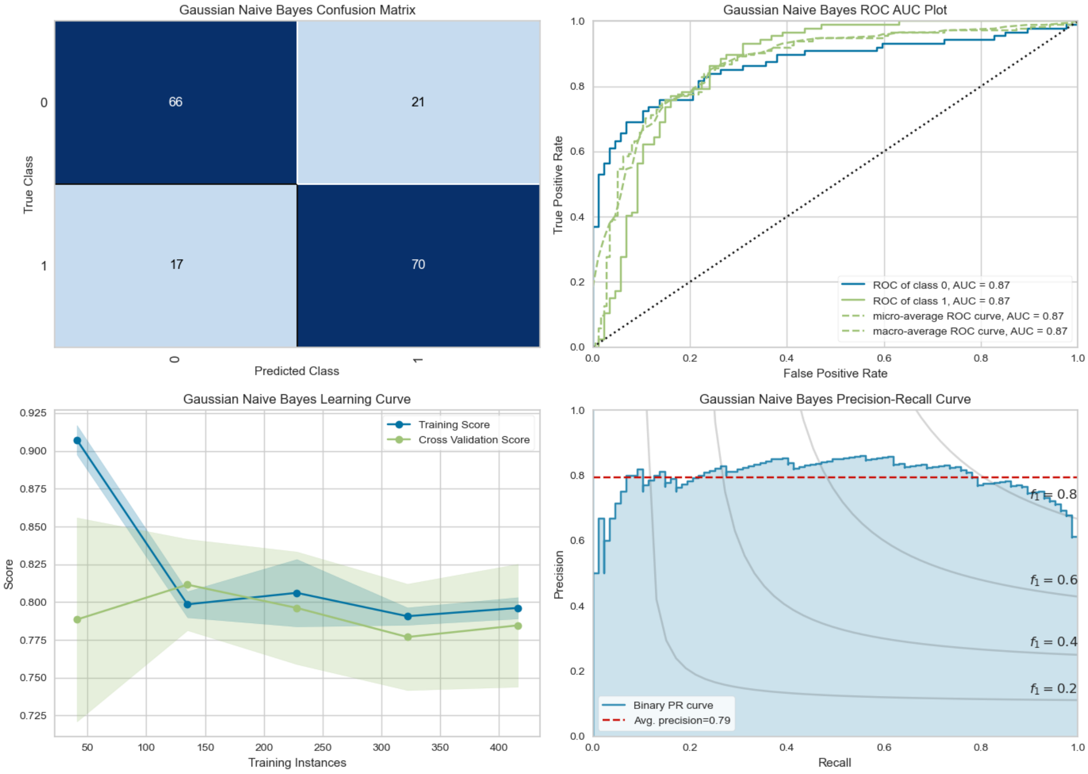


Figure 6. Gaussian Naive Bayes Algorithm Results.

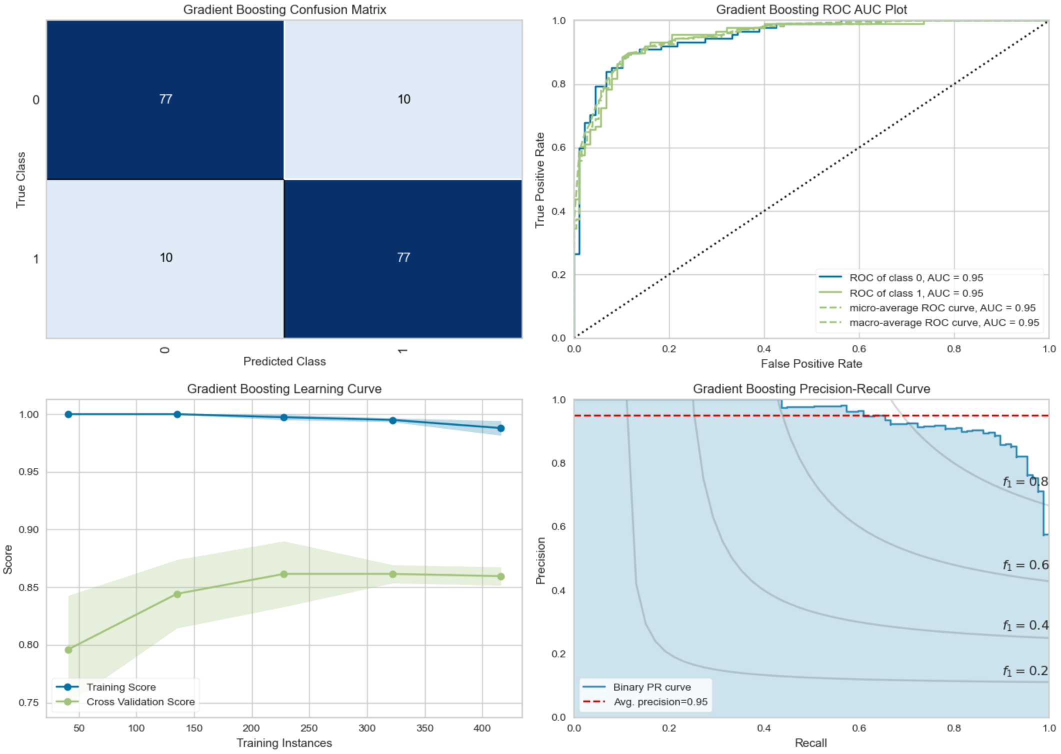


Figure 6. Gradient Boosting Algorithm Results.

Table 1. Result comperation of methods with proposed method.

Models	precision	recall	f1-score	Accuracy	Confidence Score
Gaussian Naive Bayes	0.89	0.84	0.86	0.90	0.86
Random Forest	0.91	0.86	0.89	0.89	0.88
Gradient Boosting	0.80	0.76	0.78	0.89	0.80
Linear- SVM	0.88	0.79	0.84	0.84	0.83
Logistic Regression	0.89	0.84	0.86	0.86	0.85
KNN.	0.78	0.76	0.77	0.78	0.76
Proposed Ensemble Model	**0.93**	**0.90**	**0.92**	**0.92**	**0.91**

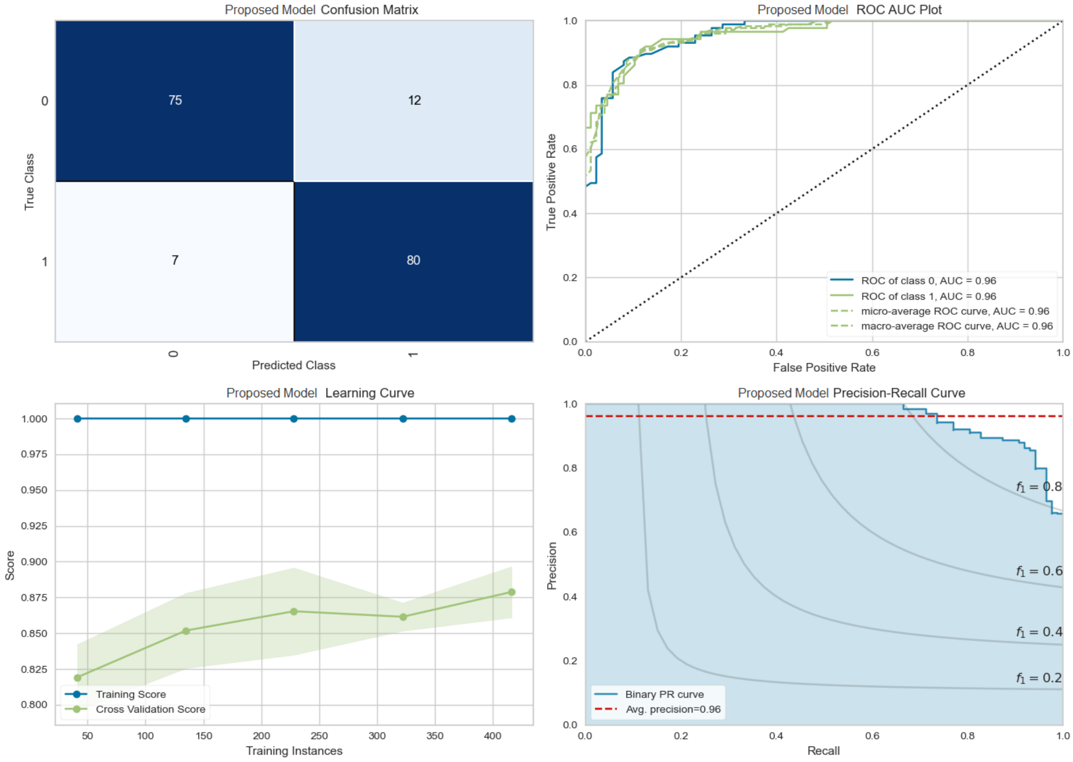


Figure 6. Proposed Model Results.

Together with a suggested ensemble model, the table compares and contrasts a number of machine learning models, such as Gaussian Naive Bayes, Random Forest, Gradient Boosting, Linear-SVM, Logistic Regression, and K-Nearest Neighbors (KNN). Accuracy, recall, f1-score, and precision are the metrics used to evaluate each model. The ensemble model performs better on all metrics by combining the advantages of the individual algorithms. It attains the maximum precision of 0.93, signifying a high percentage of correctly predicted positive outcomes. Recall of 0.90 indicates that 90% of real positive cases are successfully identified, which is important for medical predictions. The model's overall effectiveness is highlighted by the f1-score of 0.92, which displays a harmonious balance between precision and recall. Moreover, the ensemble model outperforms each individual model in accurately predicting the results, with an accuracy of 0.92. With a confidence score of 0.91, the proposed ensemble model clearly performs better than all other models when compared to Table 1's confidence scores. This high degree of confidence highlights how stable and dependable the suggested model is in making predictions. Notably, the Random Forest and Gaussian Naive Bayes models also exhibit comparatively high confidence scores of 0.88 and 0.86, respectively, demonstrating their applicability in specific situations. Nonetheless, models such as KNN (0.76) and Gradient Boosting (0.80) have lower confidence scores, indicating room for improvement in their prediction stability or the need to adjust their parameters for optimal performance. This better performance highlights how the ensemble approach can improve predictive accuracy, especially when dealing with complicated datasets like those that predict survival from oropharyngeal cancer. The success of the ensemble model highlights the benefit of combining various machine learning techniques to create a prediction system that is more dependable and robust.

4 Discussion & Conclusion

The study's comparative analysis shows how well different machine learning models predict the survival of oropharyngeal cancer. The efficacy of Gaussian Naive Bayes, Random Forest, Gradient Boosting, Linear SVM, Logistic Regression, and K-Nearest Neighbors (KNN) varies, as evidenced by their accuracy, f1-score, precision, and recall. The suggested ensemble model outperforms all single models in terms of overall performance by combining the advantages of these individual algorithms. A major advancement in the field of medical diagnostics and prognostics, this model's superior precision, recall, f1-score, and accuracy demonstrate its robustness and reliability in predicting outcomes related to survival.

The study's results underscore the noteworthy possibilities of ensemble machine learning models in the field of medical diagnostics, specifically in the area of cancer prognosis. The ensemble model is a powerful and accurate tool for predicting survival in patients with oropharyngeal cancer. It is derived from the combination of multiple individual algorithms. This method increases prediction accuracy while also offering a more thorough comprehension of the dynamics of disease. According to the findings, further research and development in the field of ensemble machine learning in healthcare should focus on exploring and improving these techniques.

References

[1] L. Wei *et al.*, "Artificial intelligence (AI) and machine learning (ML) in precision oncology: A review on enhancing discoverability through multiomics integration," *Br. J. Radiol.*, vol. 96, no. 1150, p. 20230211, 2023.

[2] J. Thagaard *et al.*, "Pitfalls in machine learning-based assessment of tumor-infiltrating lymphocytes in breast cancer: A report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer," *J. Pathol.*, vol. 260, no. 5, pp. 498–513, 2023.

[3] S.-C. Lu, C. L. Swisher, C. Chung, D. Jaffray, and C. Sidey-Gibbons, "On the importance of interpretable machine learning predictions to inform clinical decision making in oncology," *Front. Oncol.*, vol. 13, p. 1129380, 2023.

[4] W. T. Hrinivich, T. Wang, and C. Wang, "Interpretable and explainable machine learning models in oncology," *Front. Oncol.*, vol. 13, p. 1184428, 2023.

[5] M. M. Allevato, J. D. Smith, M. J. Brenner, and S. B. Chinn, "Tumor-Derived Exosomes and the Role of Liquid Biopsy in Human Papillomavirus Oropharyngeal Squamous Cell Carcinoma," *Cancer J.*, vol. 29, no. 4, pp. 230–237, 2023.

[6] J. R. Crossley, L. L. Nelson, J. Chou, and J. H. Maxwell, "Distant metastases in human papillomavirus-related oropharyngeal squamous cell carcinoma: Systematic review and meta-analysis," *Head Neck*, vol. 45, no. 1, pp. 275–282, 2023.

[7] P. A. Boot *et al.*, "Magnetic resonance imaging based radiomics prediction of Human Papillomavirus infection status and overall survival in oropharyngeal squamous cell carcinoma," *Oral Oncol.*, vol. 137, p. 106307, 2023.

[8] J. F. Mills, N. P. Monaghan, S. A. Nguyen, J. Pang, A. A. Asarkar, and C.-A. O. Nathan, "Special Issue on the Epidemiology of Human Papilloma Virus-Associated Oropharyngeal Squamous Cell Carcinoma," *Cancers*, vol. 15, no. 18, p. 4608, 2023.

[9] D. Ahn, J.-H. Kwak, G.-J. Lee, and J.-H. Sohn, "Prevalence and Characteristics of Human Papillomavirus Infection in Oropharyngeal Squamous Cell Papilloma," *Cancers*, vol. 15, no. 3, p. 810, 2023.

[10] I. Mihaylov, M. Nisheva, and D. Vassilev, "Application of machine learning models for survival prognosis in breast cancer studies," *Information*, vol. 10, no. 3, p. 93, 2019.

[11] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[12] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inform.*, vol. 2, p. 117693510600200030, 2006.

[13] W.-T. Tseng, W.-F. Chiang, S.-Y. Liu, J. Roan, and C.-N. Lin, "The application of data mining techniques to oral cancer prognosis," *J. Med. Syst.*, vol. 39, pp. 1–7, 2015.

[14] C. S. Chu, N. P. Lee, J. Adeoye, P. Thomson, and S. Choi, "Machine learning and treatment outcome prediction for oral cancer," *J. Oral Pathol. Med.*, vol. 49, no. 10, pp. 977–985, 2020.

[15] H. Patel, D. M. Vock, G. E. Marai, C. D. Fuller, A. S. Mohamed, and G. Canahuate, "Oropharyngeal cancer patient stratification using random forest based-learning over high-dimensional radiomic features," *Sci. Rep.*, vol. 11, no. 1, p. 14057, 2021.

[16] Y. Cheng *et al.*, "Using Machine Learning Algorithms to Predict Hospital Acquired Thrombocytopenia after Operation in the Intensive Care Unit: A Retrospective Cohort Study," *Diagnostics*, vol. 11, no. 9, p. 1614, 2021.

[17] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[18] P. Lambin *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, 2017.

[19] P. Lambin *et al.*, "Radiomics: extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[20] V. Kumar *et al.*, "Radiomics: the process and the challenges," *Magn. Reson. Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[21] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[22] E. M. Graboyes *et al.*, "Association of treatment delays with survival for patients with head and neck cancer: a systematic review," *JAMA Otolaryngol. Neck Surg.*, vol. 145, no. 2, pp. 166–177, 2019.

[23] J. da S. Moro, M. C. Maroneze, T. M. Ardenghi, L. M. Barin, and C. C. Danesi, "Oral and oropharyngeal cancer: epidemiology and survival analysis," *Einstein Sao Paulo*, vol. 16, 2018.

[24] M. Vedaraj, C. Anita, A. Muralidhar, V. Lavanya, K. Balasaranya, and P. Jagadeesan, "Early Prediction of Lung Cancer Using Gaussian Naive Bayes Classification Algorithm," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 6s, pp. 838–848, 2023.

[25] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," *Brief. Bioinform.*, vol. 24, no. 2, p. bbad002, 2023.

[26] S. Priya, N. Karthikeyan, and D. Palanikkumar, "Pre Screening of Cervical Cancer Through Gradient Boosting Ensemble Learning Method.," *Intell. Autom. Soft Comput.*, vol. 35, no. 3, 2023.

[27] D. Keerthana, V. Venugopal, M. K. Nath, and M. Mishra, "Hybrid convolutional neural networks with SVM classifier for classification of skin cancer," *Biomed. Eng. Adv.*, vol. 5, p. 100069, 2023.

[28] T. Shibahara *et al.*, "Deep learning generates custom-made logistic regression models for explaining how breast cancer subtypes are classified," *Plos One*, vol. 18, no. 5, p. e0286072, 2023.

[29] M. Anand, A. Velu, and P. Whig, "Prediction of Loan Behaviour with Machine Learning Models for Secure Banking," *J. Comput. Sci. Eng. JCSE*, vol. 3, no. 1, pp. 1–13, Feb. 2022, doi: 10.36596/jcse.v3i1.237.