IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 September 2023, accepted 7 November 2023, date of publication 9 November 2023,
date of current version 16 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3331720

RESEARCH ARTICLE

An Artificial Intelligence Based Approach Toward Predicting Mortality in Head and Neck Cancer Patients With Relation to Smoking and Clinical Data

NAMAN DHARIWAL, RITHVIK HARIPRASAD, AND L. MOHANA SUNDARI
Department of Software Systems, Vellore Institute of Technology, Vellore 632014, India
Corresponding author: L. Mohana Sundari (mohanasundari.l@vit.ac.in)

ABSTRACT Head and neck cancers are one of the most common cancers in the world which affects the mouth, throat, and tongue regions of the human body. Lifestyle factors such as smoking, and tobacco have been long associated with the generation of cancerous cells in the body. This paper is a novel approach towards extracting the correlation between these life factors and head and neck cancers, supported by crucial cancer attributes like the tumor-node-metastasis and human papilloma virus. Mortality prediction algorithms in cases of head and neck cancers will help doctors pre-determine the factors that are most crucial and help deliver specialized and targeted treatments. The paper used eight machine learning and four deep learning hyper-parameter tuned models to predict the mortality rate associated with head and neck cancer. The maximum accuracy of 98.8% was achieved by the gradient boosting algorithm in the paper. The feature importance of smoking and human papilloma virus positivity using the same classifier was approximately 4% and 2.5% respectively. The most influential factor in mortality prediction was the duration of follow-up from diagnosis to the last contact date, with 40.8% importance. Quantitative results from the area under the receiver operating characteristic curve substantiate the classifiers' performance, with a maximum value of 0.99 for gradient boosting. This paper is bound to impact many medical professionals by helping them predict the mortality of cancer patients and aid appropriate treatments.

INDEX TERMS Head and neck cancer, human papillomavirus, machine learning, smoking, deep learning.

I. INTRODUCTION

Cancers are one of the biggest health threats in the world, affecting millions each year. Cancers are diseases which affect the cell structure of the body, causing cells to grow uncontrollably resulting in a spread to other parts of the body. These surplus cells collect at different places of the body resulting in formation of tumors, which are just a lump of excess tissue. These tumors can in turn be malignant or benign by the process called metastasis, the malignant or fetal tumors invade the other parts of the body and travel to organs creating more potentially deadly tumors in the body. Therefore, a person is always at risk of recurrence of cancer when a malignant tissue is removed due to the possibility of cancerous cells still being present in the body. The benign cells on the other hand don't invade any nearby tissue structure and once removed do not recur in most cases. According to the National Cancer Institute (NIH:NCI) [1] breast cancer, lung cancer and thyroid cancer are among the topmost common occurring cancers (in 2020). They also state that approximately 39.5% of the population will be fighting cancer at some point in their lives. By the year 2020, the number of new cancer patients annually will reach 29.5 million and the annual death toll will reach 16.4 million. According to [2] there were approximately 18 million new cancer cases in 2020 alone. The paper also revealed that approximately

The associate editor coordinating the review of this manuscript and approving it for publication was Longzhi Yang.

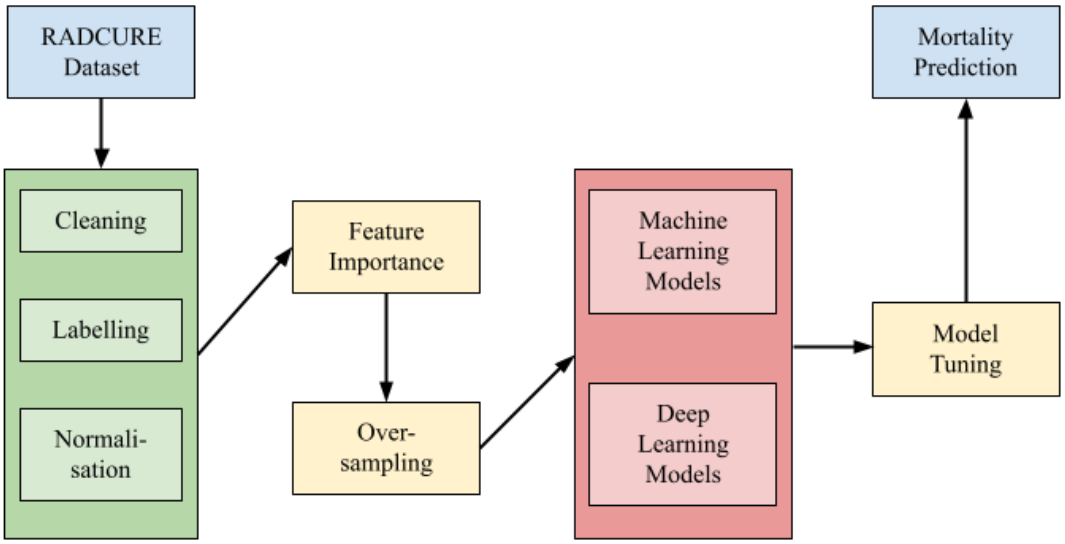

FIGURE 1. Flow chart representing the methodology of the paper.

10 million people died due to cancer complications in the same year. About 33% of all the cancer cases reported were related to smoking tobacco.

This paper focuses on the cancer of the head and neck, also known as squamous cell carcinomas of head and neck. These cancers usually begin in the squamous cells that are present in the linings of the mucosal regions like the inside of the mouth, voice-box and throat. Head and neck cancers are broadly formed in the following five regions [1]:

1. Oral cavity- This consists of the lips, cheek lining, gums, frontal two thirds of the tongue, the roof of the mouth, the bottom of the mouth below the tongue and the sections behind the wisdom teeth.
2. Pharynx- The throat is the hollow tube which goes from the esophagus to behind the nose. It is approximately 5 inches in length and consists of 3 sections, the nasopharynx, the oropharynx, and the hypopharynx.
3. Larynx- It basically consists of the region made of cartilage, just below the pharynx, and the epiglottis. It is known commonly as the voice box.
4. Paranasal sinuses- These are the small hollow cavities in the bone surrounding the nasal cavity.
5. Salivary glands- These are the glands that are present near the jawbone at the base of the mouth and are responsible for creation of saliva in the mouth.

The NIH:NCI [1] presents an elaborate study on the possible causes of head and neck cancers. As discussed earlier too, the use of tobacco (smoking, chewing, or snuffing) and alcohol (both together) remains one of the most influential factors in causing generation of squamous cell carcinomas in the regions like the oral cavity, hypopharynx, and the larynx. The consumption of 'paan' or betel quid is also a major cause behind mouth cancers. In three quarters of the total carcinoma cases, being infected with human papillomavirus (HPV) type 16 causes cancerous cells to develop in the tonsils and under-tongue regions. Entry of foreign particles like wood powder, glass powder, very minute fragments metal and asbestos, from occupation sites, may cause the nasal cavities to be affected fatally and carcinogenic. Radiation has been long proven to be carcinogenic in humans and it may lead to oral cancers too. Ancestry and genetics also affect the chances of a person having cancer. Asian ancestry and genetic disorders like fanconi anemia are some of the leading causes.

Deschler et al. [7] in association with American Academy of Otolaryngology- Head and Neck Surgery Foundation present the details about identifying the stages of head and neck cancers using the TNM or the tumor-node-metastasis staging system. T gives the primary characteristics of the tumor like size and/or location. N represents the degree of regional lymph node involvement. M gives the information about the presence or absence of distant metastasis of the cancer. TNM system has different representative values respectively for different types of cancers. The possible combinations of these T, N and M values then result in identification of the cancer stage from 0 (least harmful) to VI (most harmful).

Head and neck cancer present a wide range of symptoms that serve as potential indicators of its presence. Localized discomfort, often persistent and constant, can manifest in areas such as mouth, throat or ears, hinting at underlying malignancies [3]. A noticeable alteration in voice quality, marked by hoarseness or rough tone could indicate laryngeal engagement, where cancerous growth may have affected normal functioning of vocal cords. Unexplained reduction in body mass and weight loss could be indicative of the cancer's metabolic impact. Head and neck cancer can also trigger pain in areas that may seem unrelated, suggesting a phenomenon

known as referred pain. The emergence of a lump or growth in the neck area is often a sign indicating potential tumor spreading to lymph nodes or surrounding tissues. Hemorrhaging and experiencing difficulty swallowing, known as dysphagia, are also additional symptoms that should not be ignored.

Timely recognition and medical attention for these indicators can make the pivotal difference between life and death. Regular medical check-ups, particularly of risk factors like tobacco and alcohol use are present, can aid in spotting these symptoms early. If a diagnosis is made, treatment options include radiation, chemotherapy, and surgery [14]. Radiotherapy is particularly effective in early stages and is either the main course of treatment or alongside other therapies. Advances in imaging and radiation delivery have revolutionized this treatment approach. Various advanced methods like tomotherapy, stereotactic radiosurgery and proton therapy have been introduced although many require validation through clinical trials. Chemotherapy has evolved from being just a method of palliative care to a key element in curative treatments for advanced stages. Targeting epidermal growth factor receptor (EGFR) has emerged as a promising approach. Novel strategies involve combining EGFR inhibitors with other targeted agents or traditional treatments like chemotherapy and radiotherapy. Surgery is a standard treatment for head and neck cancer often constrained by tumor size and the goal of preserving organ function. This approach is recommended with chemoradiotherapy when residual disease is suspected.

A research [4] shows that patients identified to earlier stages showed better prognosis as compared to those of later stages. In the initial stages, there are choices such as radiation therapy that are more suitable [5]. As the stages progress, these treatment options become less viable and much drastic treatments such as surgical removal of the affected area which come at a high cost, both financially and physically. After this the patient must undergo speech and swallowing therapy to ease back into a normal life. The patient is compelled to undergo successive follow-up procedures in order to detect any recurrence of the tumor [6]. These follow ups cause the patient unnecessary distress and morbidity and can have significant cost implications for the healthcare system.

This paper aims to find the correlation between mortality in head and neck cancers with lifestyle factors like smoking. Standard cancer indicators like TNM and HPV-positivity were also used to add support to the paper. The 12 different prediction models were hyper-trained by machine learning (ML) and deep learning (DL) techniques to predict mortality. The trends in mortality and the importance of lifestyle choices in head and neck cancer patients are aimed to be identified by extracting features importance. The aim is to assist doctors in predicting mortality in cancer cases and providing optimized treatment to prevent loss of life.

II. LITERATURE REVIEW

Successful use of ML and DL techniques for prediction of malignancy of cancer or extracting important features from lab image results has long been proven. Kourou et al. [8] present a review of the various supervised machine learning techniques like artificial neural networks (ANN), bayesian networks (BN), support vector machines (SVM) and decision trees (DT) that help in cancer prognosis and prediction. They reveal that SVM and ANN are the most widely used machine learning techniques when it comes to predicting the outcome of cancer patients, mainly due to their high accuracy scores. However, they also reveal that choosing the most appropriate machine learning model depends on various factors like the data and expected outcomes. In another independent research by Murthy et al. [9] the need for early prognosis of cancer for effective and accurate treatment is justified. In addition to the aforementioned machine learning techniques, the research presents views on Recurrent Neural Network (RNN) and Deep Neural Network (DNN) methods, further stating that due to these technologies the rate in prediction has gone up to 99.89% in some cases. The paper talks in great depth about several papers published on predicting cancer results from many datasets related to oral cancer, breast cancer, and many more. They agree that machine learning techniques are in fact essential in cancer diagnosis with the help of adequate datasets. However, the researchers also correctly identify the limitations of using machine learning techniques for cancer prediction, they state that there must be bias free datasets and with the in-flow of high-quality data, development of better high-quality models is necessary.

In a research, Dong et al. [15] compared the efficacy of three models for predicting survival of patients of oral cancer. One a deep learning-based survival prediction algorithm named DeepSurv, the second a Random Survival Forest (RSF) and lastly, a Cox Proportional Hazard model (CPH). They also performed statistical analysis using various methods namely, Mann-Whitney U test, Chi square test, Fisher's exact test, and Cochran-Armitage Trend test. The models were trained using different number of features in each run to see the effect of the number of variables on the models. They started with the five most important features and consequently added the remaining statistically insignificant features one by one. They observed that with the increase in the number of features, the c-index of DeepSurv rose while that of CPH and RSF decreased. Overall, the DeepSurv showed the highest c-index of 0.810 and 0.781 for the training and testing set.

The researchers in another paper [16], evaluated the effects of Convolutional Neural Network (CNN) on Computed Tomography (CT) image data of head and neck cancer records. They extract radiomics features from the CT images by using the slice that contains the greatest number of tumors. These radiomics features are used in the model training process. To better the results, André et al. implemented cross validation using K-folds where the number of folds were taken to be 5. Finally, the most robust network achieved an AUC of 0.88 in predicting distant metastasis. The researchers further stated that the CNN model removes the necessity for intricate feature engineering which is a common requirement in traditional radiomics. This method empowers the

TABLE 1. Machine learning and deep learning model results.

S. No.	Classifier	Accuracy	Precision	Recall	F1-Score
Machine Learning Models					
1	Logistic Regression	0.727019	0.726764	0.727019	0.726587
2	Decision Tree	0.987001	0.98735	0 .987001	0.987008
3	Random Forest	0.87558	0.875558	0.87558	0.875521
4	Gradient Boosting	0.988858	0.989115	0.988858	0.988864
5	Support Vector Machine	0.780873	0.781055	0.780873	0.780353
6	K- Nearest Neighbors	0.84494	0.857912	0.84494	0.844472
7	XGBoost	0.895079	0.895551	0.895079	0.895141
8	LightGBM	0.806871	0.807031	0.806871	0.806926
Deep Learning Models					
9	Convolutional Neural Network	0.917005	0.950676	0.885209	0.916613
10	Deep Neural Network	0.874997	0.909137	0.843418	0.874651
11	Recurrent Neural Network	0.89725	0.939858	0.857215	0.896245
12	Artificial Neural Network	0.883143	0.924137	0.842215	0.880739

algorithm to independently detect relevant image details, potentially resulting in enhanced prediction accuracy.

Kazmierski et al. [10] very recently presented a paper on the prognostic modeling in head and neck cancer and evaluating the impact using deep learning techniques and radiomics. The researchers analyzed the performance of 12 different deep learning models, such as fuzzy logistic regression, on an institutional dataset of 2552 head and neck cancers patients' electronic medical records and pretreatment radiological images. The various machine learning model development and accuracy analyses were assigned to independent experimentalists by the authors. The multitask learning model performed the best in terms of accuracy on the clinical data. It also achieved high accuracy in prognosis for lifetime survival prediction. Three benchmark models were created by the authors for the generation of baseline comparison points. The categorical data was encoded using one-hot encoding. Further the data was subjected to feature selection and the models' hyperparameters were tuned using grid search techniques and K-folds validation (K was taken as 5). The three radiomics, volume and clinical models were compared on the basis of area under the curve (AUC) and clinical performance was the best among the three with AUC=0.74 as compared to AUC=0.71 in other cases. The authors signify that the combination of ML and simple prognostic factors outperformed several complex CT radiomics and deep learning approaches. However, like other researchers, they warn that although ML models offer a variety of prognostic possibilities for HNC patients, their prognostic utility is influenced by variances in patient groups and need further validation. The researchers also admit that other machine learning ensemble learning techniques like BN may perform better in predictions.

Doctors and researchers have, in the past, established a strong link between smoking and cancers, especially oral and lung cancers [11], [12], [13]. Researchers Lee et al. [11] present a paper on risk prediction models that can and were used for head and neck cancers. They used various lifestyle predictors like smoking, consumption of alcohol, education status of the individual, etc. The results of the prediction were separated by age in most cases with support of other attributes. A 45-year-old woman who had completed high school, smoked more than 20 cigarettes per day for more than 20 years, drank three or more alcoholic beverages per day had a 20-year absolute risk of head and neck cancers of 4.77%. For males with comparable features, the risk was 5.72% during a 20-year period. Their statistical results clearly justify that smoking and alcohol are in fact highly influential features in any head and neck cancer dataset. Mirestean et al. [12] also presented a paper on the use of radiomic machine learning for head and neck oncology. They identify that oropharyngeal squamous cell carcinoma has a higher incidence than other head and neck cancers due to HPV infections, which are biologically and clinically distinct from HPV-negative oropharyngeal cancer, which is frequently linked to smoking and alcohol use. A major result discussed in the research stated that Positive HPV illness is more responsive to radiation treatment and chemotherapy and is associated with much greater rates of healing when compared to HPV negative head and neck cancer, which is typically brought on by smoking and alcohol use. Therefore, the research clearly indicates that lifestyle factors and personal habits play a significant role in the prediction and treatment of cancer. Another research by authors Gritz et al. [13] in 1993 was focused on the long-term smoking cessation in head and neck cancers. In the paper

TABLE 2. Feature importance using Gradient Boosting Classifier.

Feature	Importance Percentage
Length FU	40.87
Age	22.36
T	5.99
N	4.94
Ds Site	4.73
Stage	3.82
Smoking Status	3.81
Dose	3.07
HPV_Yes, positive	2.53
Fx	2.19
Sex	1.52
HPV_Not tested	1.34
Tx Modality_RT alone	0.81
HPV_Yes, Negative	0.69
Chemo?	0.67
Tx Modality_RT + EGFRI	0.3
Tx Modality_ChemoRT	0.28
Smoking PY	0.08
M	0.01
Tx Modality_Postop RT alone	0.01

the participants were 186 individuals with newly discovered initial primary squamous cell carcinomas of the upper aerodigestive tract and recent cigarette smokers. Their paper found serious implications of smoking on patients' health and the researchers urged patients with head and neck cancer to stop smoking in a methodical, quick manner.

A research [17] by Carole et al. depicts that HPV infections are a significant contributor to head and neck squamous cell cancers, particularly in the oropharynx. The research also shows that the past 3-4 decades have shown a decrease in risk factors such as alcohol and tobacco use but still show an increase in oropharyngeal cancer. It has been found that the HPV positive patients tend to be younger and more likely to be male. Ongoing research focuses on targeted therapies, immune therapies and HPV targeted vaccines that may reduce the chances of cancer if HPV is caught and treated at its early onset.

After reviewing several literatures related to head and neck cancer and its predictions models, gaps were found in the studies. It is seen that though smoking is studied as a crucial factor in predicting cancer in general, it is not yet related to the mortality trends in head and neck cancer patients.

TABLE 3. Mean AUC Values from the ROC curves of the 10 classifiers.

Classifier	Mean AUC
Logistic Regression	0.81
Decision Tree	0.95
Random Forest	0.92
Gradient Boosting	0.99
K- Nearest Neighbors	0.9
XGBoost	0.94
Convolutional Neural Network	0.93
Deep Neural Network	0.9
Recurrent Neural Network	0.91
Artificial Neural Network	0.91

On the other hand, cancer-based prediction studies follow the trend of using CT imaging datasets rather than clinical data which provides more dimensions to the actual condition of the patient. Clinical data possess inputs on various lifestyle choices, demographic factors and treatment facilities and outcomes that are absent in imaging datasets. Therefore, in this paper, the clinical data related to head and neck cancer is used, thus, offering a unique perspective towards mortality predictions.

III. METHODOLOGY

For ease of operation the methodology is divided in six segments as described in the following sections. A flow chart representation is presented via figure 1.

A. DATASET DESCRIPTION

The data used in this paper was sourced from the publicly available Computed Tomography Images from Large Head and Neck Cohort RADCURE dataset [18]. The RADCURE dataset is an extensive collection of CT images and the correlated clinical data from 3346 head and neck cancer patients treated with definitive radiation therapy at the University Health Network in Toronto, Canada, collected from 2005-2017.

The clinical data comprises extensive details for each patient covering demographic, clinical, and treatment aspects. The dataset comprises a comprehensive array of columns presented in both numeric and categorical format, capturing essential patient and clinical information. For instance, while numeric columns such as age and dose are present, categorical columns like sex, site and subsite contribute to the diversity of information captured. These columns encompass patient identification, demographic details (age, sex), functional status (ECOG PS), smoking history, cancer attributes (site, subsite, stage), treatment specifics (modality, chemo administration), radiotherapy details (start date, dose, fractions, technology), follow-up

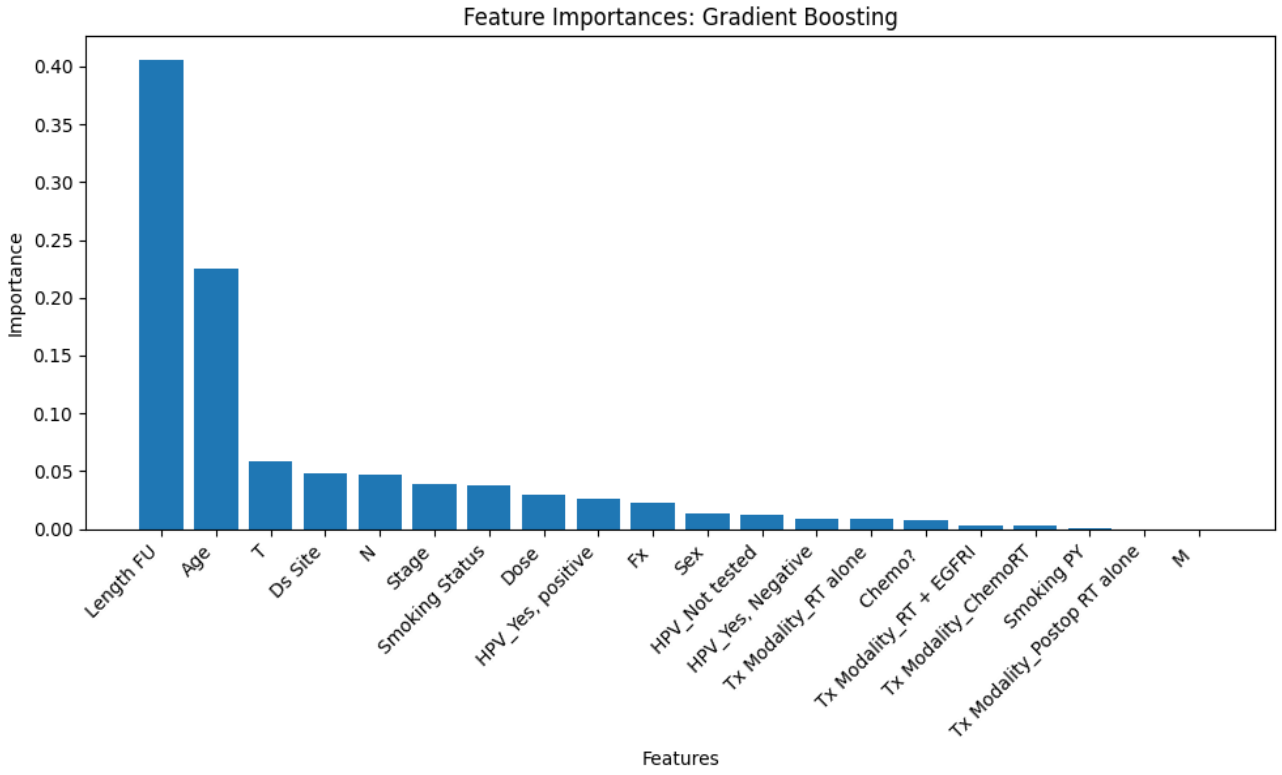

FIGURE 2. Feature importance using gradient boosting.

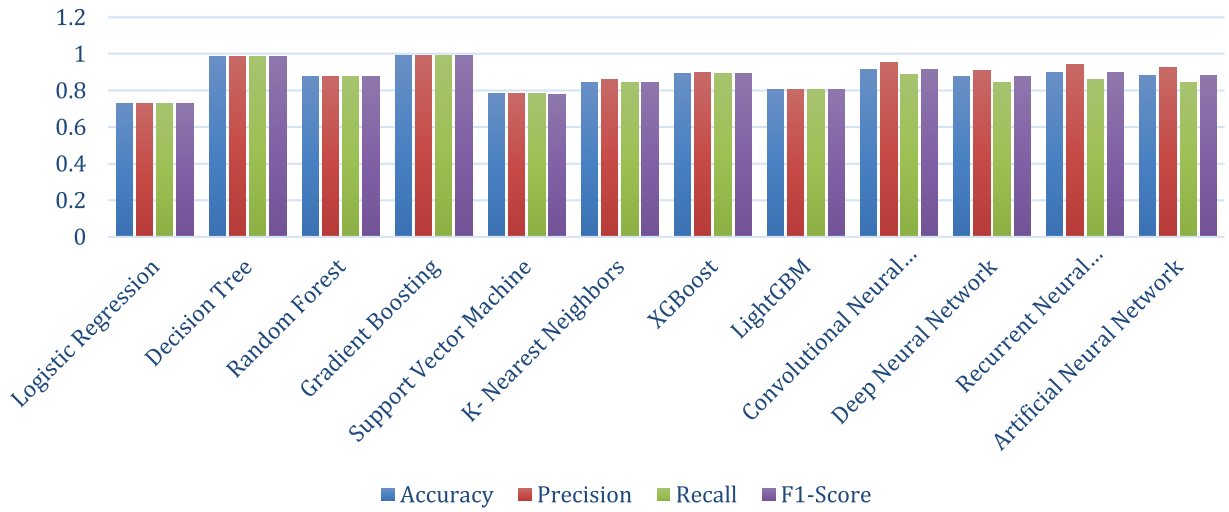

FIGURE 3. Bar graph representation of machine and deep learning models.

information (last contact, vital status, duration), and clinical outcomes (date and cause of death, relapses, second cancer diagnosis).

For the purpose of this paper only the clinical data has been used as it provides essential contextual information about patients including demographic details, medical history, treatment approaches and the outcomes. Clinical data is also critical in identifying risk factors that might not be apparent from imaging.

B. DATA PRE-PROCESSING

The dataset obtained had various imperfections like missing data values, duplicate records, and probable outliers. Data pre-processing is performed, involving multiple steps

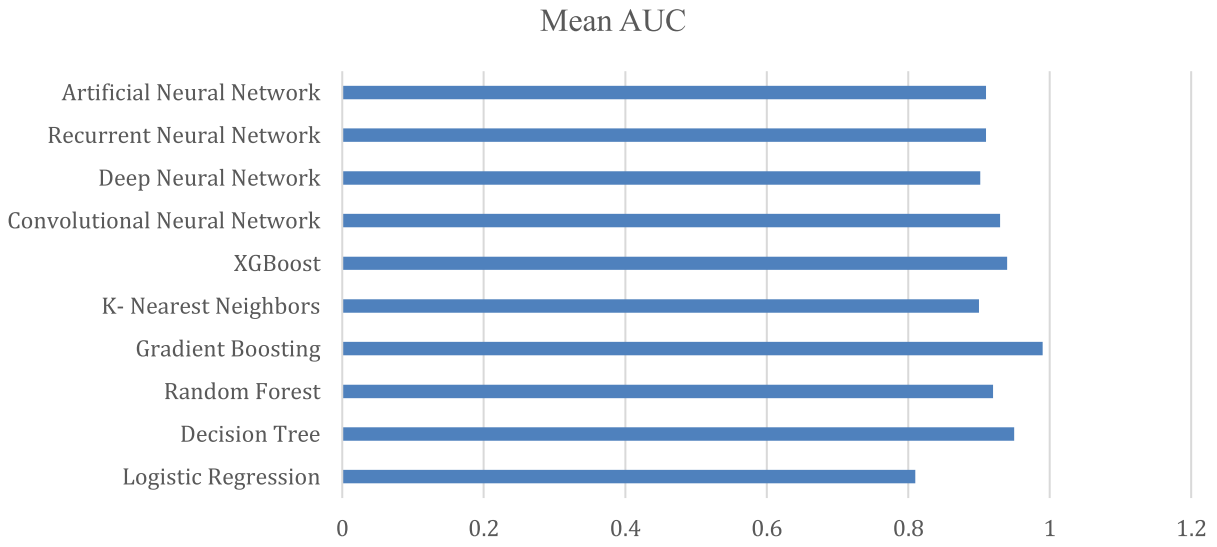

FIGURE 4. Comparative visualization of AUC from the ROC curves.

to prepare the raw dataset into processed, organized, and systematic data for eventual training using machine learning and deep learning models. The pre-processing was divided broadly in the following 3 stages:

1) CLEANING

The first stage of data pre-processing was the data cleaning stage. The paper uses the python library pandas and its pre-defined methods like '*isnull()*' and '*describe()*' to carefully inspect the data. The dataset did contain missing values in several columns spread across many rows, however, no data duplication was identified. To handle the missing data, data imputation was performed, filling the missing numerical values in a numerical column with the mean of the existing values. However, in the case of categorical data, the median of the column was used as the new value in missing cells. If any column had an extremely large number of missing values, then that column was deleted due to potential bias or other problems being generated during training. The irrelevant columns such as patient id and date-time were also deleted. The data was refined from deleting the rows. It is to be noted that further column deletion was performed after performing feature importance and the columns with zero relevance were removed for easy computation.

2) LABELING

The dataset contained both numerical and categorical data columns. Therefore, 3 techniques were considered to convert the categorical data into respective numeric data. The first labeling technique was one-hot encoding which created a separate binary column for each unique value in a categorical column. The result was a dataset with 121 columns, which posed as a drawback due to increase in volume and dimensionality of data. The second technique was the use of label encoding technique which assigned a random integer to each of the unique values in a categorical column. The resulting dataset, though retained the original dimensionality, induces disproportionate relation between values of two columns. Therefore, the final methodology used was to manually encode each unique value of the columns of the dataset. For columns like gender, label encoding was used, however, in columns like T, N, M, smoking status etc. Manual encoding was preferred.

3) NORMALIZATION

After the successful completion of the previous stages, a numerical full dataset with no missing values was obtained. However, one big challenge remaining was the large ranges of values per column. These large fluctuations would eventually cause computational delay and therefore, data normalization was performed. All the numerical data in each column was converted to a standardized range to ensure no single feature dominated during model training and testing. The '*StandardScaler()*' method from the sklearn library was employed. This particular method was chosen because of its ability to normalize the data into a range such that the overall mean becomes 0 and standard deviation becomes 1, thus in turn avoiding scale bias in the models. Another important reason for choosing the method was its capability to handle outliers efficiently and robustly.

C. FEATURE EXTRACTION

Medical data contain many columns that lead to high dimensionality that further leads to low accuracy in classification models [19]. Feature extraction is therefore critical in reducing the dimensionality of the data by only keeping necessary columns and removing the rest. This process improves computational efficiency, accelerating model training, decision-making, and enhances accuracy by focusing on only the important data aspects. Both the random forest classifier and the XGBoost classifier employ decision trees as their foundational components. However, their learning

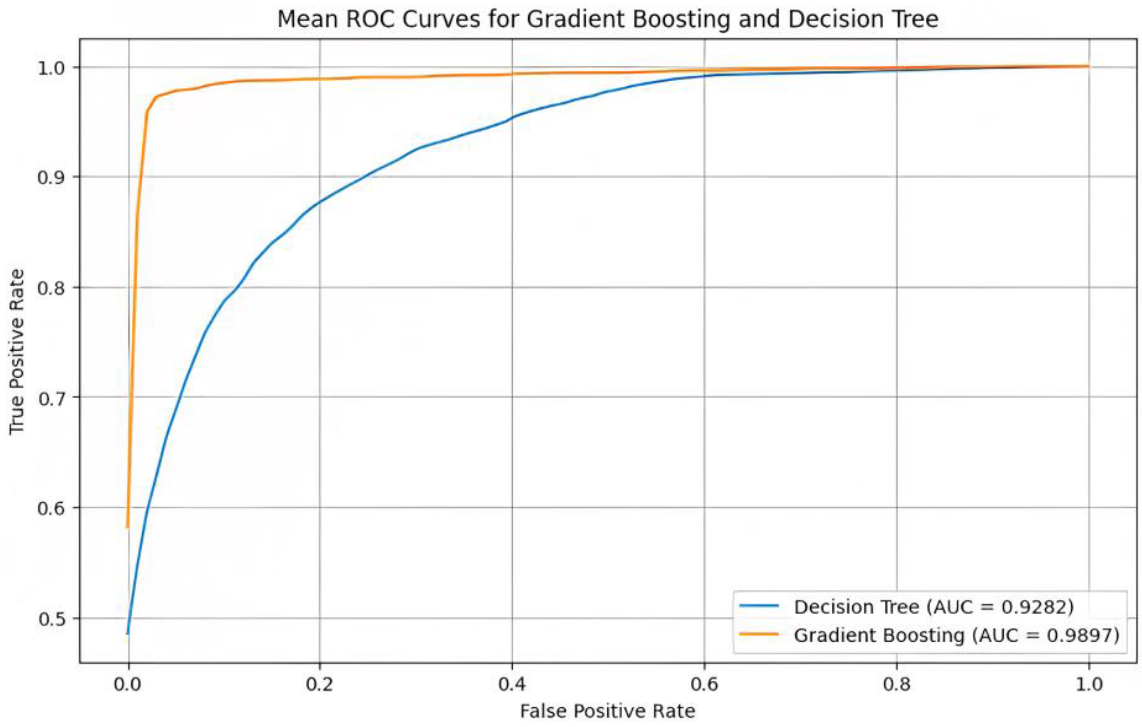

FIGURE 5. The ROC curves for the gradient boosting and decision tree algorithms.

techniques differ, the random forest classifier utilizes Bagging, while the XGBoost classifier employs Boosting [20]. The '*feature_importances_*' attribute was utilized after fitting the data into models, providing the feature importance for each feature, respectively. This data was then used to remove the features that were found to have low importance therefore decreasing the dimensionality of the data.

D. OVERSAMPLING

Class imbalance happens when, in a binary classification issue, one class (the minority class) has considerably fewer occurrences than another class (the majority class). This disparity might result in biased and erroneous model performance. Tests such as class distribution ratio computation were performed to identify any imbalances in the dataset. It was observed that the dataset was imbalanced and subsequently subjected to oversampling techniques. The target attribute, mortality status, contained two classes, namely dead and alive, in which the dead data was in less proportion to the alive data. To balance the dataset, random oversampling was employed by duplicating some of the data points at random from the minority class. The process was conducted with high supervision to avoid any chances of overfitting and loss of information.

E. TRAINING MODELS

After completing all necessary transformations on the dataset, attention was turned towards training models for comparison based on metrics such as accuracy, f1-score, precision, and recall values. Traditional ML models, modern ML models, and DL models were employed for the paper.

1) MACHINE LEARNING MODELS

In this paper, traditional machine learning algorithms such as Logistic Regression, Random Forest, Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, XGBoost, and Gradient Boosting were used. These models are classification models used to categorize data into predefined categories or classes. In this case, the classification is between 'dead' or 'alive'. The widely used Logistic regression serves as a classification technique, yielding a binary outcome that predicts the probability of an event occurring (either 0 or 1) based on input variables [21]. Decision Tree is a supervised machine learning method used for classification and regression tasks by iteratively splitting data based on specific criteria. It employs nodes for data splitting and leaves for decisions. Random Forest is when there is a collection of decision trees. It is an ensemble learning technique that builds multiple decision trees during training and combines their predictions to make more accurate and robust predictions. Naïve Bayes is a simple algorithm relying on conditional probability. It utilizes a probability table as its model, updating it with training data. This table, based on feature values, provides class probabilities for predictions. In SVM, the classification is done by first finding the hyperplane which is a decision boundary between different classes

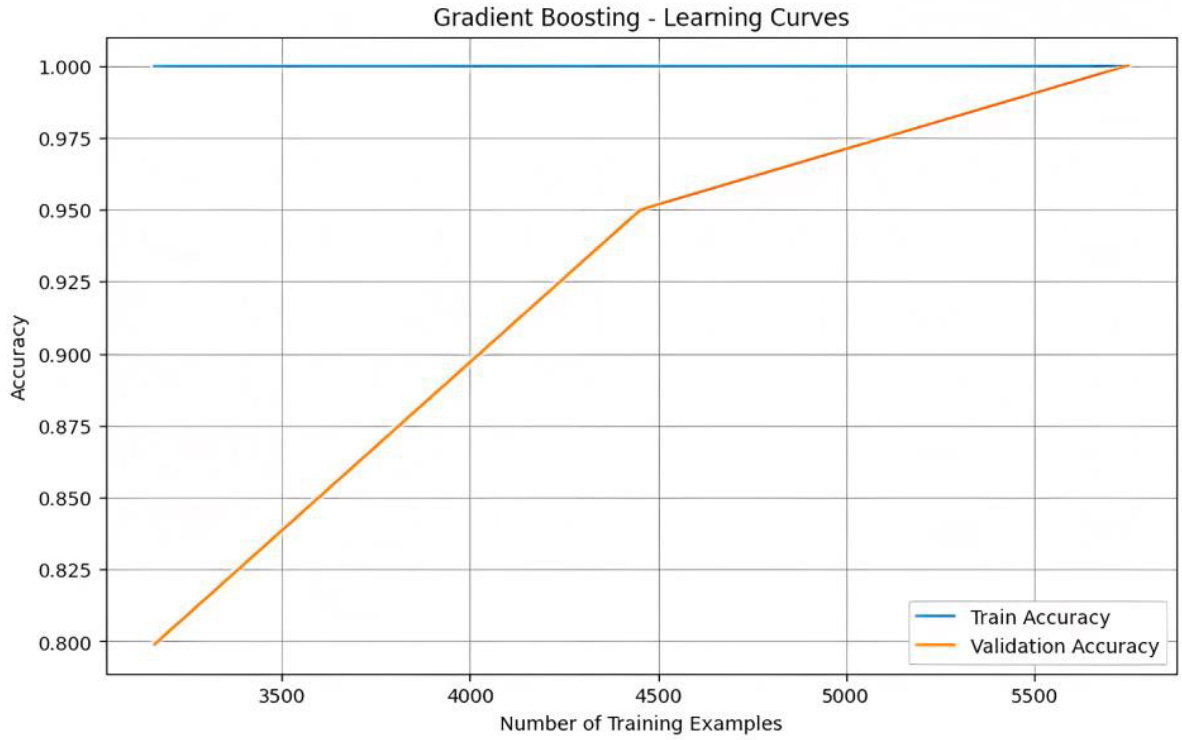

FIGURE 6. The learning curves for the gradient boosting algorithm.

of data points. The hyperplane is crucial in separating data points of different classes but may be less effective when data cannot be linearly separable. The KNN Algorithm is a classification method utilizing a database of grouped data points for classification tasks. KNN is non-parametric, not assuming any underlying data distribution. Gradient Boosting is an iterative machine learning algorithm that assembles weak learners, which are models that are slightly better than random guessing, into a robust and accurate predictive model. XGBoost is a Classifier that uses decision trees and gradient boosting thereby increasing the training speed of the decision trees by reducing computational complexity [22].

2) DEEP LEARNING MODELS

Four DL models: CNN, DNN, RNN and ANN were also used. All the four models work on the concept of layers: input, hidden and output. Each of the four models were chosen on the basis of the advantages they provide in predictions. CNNs, though popularly used in image classification, provide excellent results in textual datasets as well due to its high accuracy in feature extraction and pattern recognition. DNN are furthermore effective in accurately predicting results by extraction of important features and thus, were chosen for the paper. RNNs are best to study temporal dependencies present in data and thus, provide another angle to the prediction study. ANNs are one of the most traditional DL models and it is only appropriate to compare the modern models with the accuracy of the traditional model, which is considered to be one of the best in textual prediction scenarios. The study uses pre-defined neural network architectures presented by Tensorflow, Keras and PyTorch in a python environment as they provide high quality APIs that allow training and evaluating these models.

F. MODEL TUNING

The mentioned ML and DL models initially yielded satisfactory results. The goal was to enhance their performance and evaluation metrics through hyperparameter tuning and K-fold operations on the training models. K-folds, a cross-validation technique used to mitigate problems related to overfitting, was employed. The dataset is thus split into 'K' sections or splits and processed to obtain best performance and accuracy. One crucial reason behind performing K-folds was to eliminate any chance of overfitting that might have occurred due to the random oversampling process in the oversampling stage. Furthermore, the sklearn library is utilized, and the '*GridSearchCV*' function is imported. This technique is employed to identify the best possible permutations of the hyperparameters, resulting in improved accuracy and performance. Further hyper parameter tuning was performed on Dl models like the tuning of '*epochs*', '*batch_size*', '*random_state*', etc. In the case of ML models, parameters like '*max_depth*', '*max_iter*' and many more were tuned to get the most optimized results in the generalized models. Throughout the tuning process, various parameter values were experimented with, and the

most accurate values were determined through comparison among them.

IV. RESULTS AND DISCUSSIONS

In this paper, the aim was to predict mortality rates in relation to lifestyle choices such as smoking and other standard attributes studied in cancer research. By employing the previously mentioned methodology, accuracy, recall, precision, and f1-score metrics were achieved for the 12 artificial intelligence architectures, as depicted in Table 1 and visualized in figure 3.

As observed, the maximum accuracy of 98.88% was observed using the gradient boosting model. Another model, DT classifier also achieved a close accuracy score of 98.70%. In most cases, it was observed that machine learning models demonstrated higher accuracies in predicting mortality in head and neck cancer compared to deep learning models. The best performer of the DL models was the CNN that had an accuracy of 91.70%. The learning curve of gradient boosting algorithm is shown in figure 6.

The importance of the features involved in the prediction models was also extracted. Table 2 presents the feature importance results for the most accurate training model, gradient boosting. The same is visualized in Figure 2. It was observed that the attribute 'length FU,' representing the duration of follow-up from diagnosis to the last contact date, had the most significant impact on the predictions. This is justified as shorter follow-up duration signifies mortality where longer durations signify continuations in treatments and prolonged life. The second most important feature is the age of the patient, the older the patient the higher the risk of mortality. These trends are followed by T and N factors which signify the size and lymph node involvement and spread; thus, they are treated as crucial factors in any cancer prediction. '*Ds Site*' represents the primary cancer site, thus having influence on mortality. The paper aimed to determine the correlation of 'Smoking Status' as a lifestyle factor. It is the 6th most important feature involved in the prediction of mortality after the more traditional aforementioned attributes. Thus, smoking plays a very crucial role in head and neck cancers and can be used as a feature in mortality predictions. The following two attributes with high importance are '*stage*' and '*dose*', which represent the stage of cancer and the dosage of radiotherapy. It is substantial as the higher the stage of cancer a patient has the higher their risk of death. '*HPV_Yes, positive*' was the 9th most influential attribute in prediction. These results are in-line with the observations from section II. Hence, this paper clearly demonstrates the significance of lifestyle choices in relation to mortality associated with head and neck cancers.

Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) are some of the performance measures for the binary classifiers. Table 3 shows a comprehension of the mean AUC outcomes from the ROC curves of the classifiers. It is observed that gradient boosting, the most accurate model, has an AUC value of approximately 1. Thus, showing the quality of performance. The results are visualized in figure 4. The ROC curves, used to calculate the AUC values, for the top two most accurate models, gradient boosting, and decision trees, are shown in figure 5.

Throughout the paper's progression, a few limitations in the dataset used for this research were identified. The absence of attributes specifically related to tobacco use, one of the major contributing factors to the occurrence of mouth and neck cancers, was observed. Also, the existing attribute '*M*' which refers to metastasis, had improper data and influenced the mortality prediction by a negligible percentage. This is in stark contrast to the various other studies conducted that suggest the impact of metastasis in cancers. However, these limitations posed no contradiction to the aforementioned results.

V. CONCLUSION AND FUTURE WORKS

In this paper, a novel approach was presented for predicting mortality in head and neck cancer patients in relation to lifestyle choices like smoking. The gaps in the existing literature were recognized, and the potential contribution in the ongoing battle against cancer was identified. Multiple training models were developed and compared, resulting in high accuracy. Various concepts of artificial intelligence were utilized to enhance the accuracy and precision of predictions. The goal is to aid medical professionals and doctors in improving cancer treatment and minimizing mortality rates as much as possible. Future research aims are to incorporate the models with other pre-existing models to create a unified cancer mortality prediction application that will help doctors around the world predict and treat cancer effectively.

REFERENCES

[1] National Cancer Institute. *Comprehensive Cancer Information*. [Online]. Available: https://www.cancer.gov
[2] *Cancer Research UK*. [Online]. Available: https://www.cancerresearchuk.org
[3] R. W. Dolan, C. W. Vaughan, and F. Nabil, "Symptoms in early head and neck cancer: An inadequate indicator," *Otolaryngol. Head Neck Surgery*, vol. 119, no. 5, pp. 463–467, Nov. 1998.
[4] H. Lim, D. H. Kim, H.-Y. Jung, E. J. Gong, H. K. Na, J. Y. Ahn, M.-Y. Kim, J. H. Lee, K.-S. Choi, K. D. Choi, H. J. Song, G. H. Lee, and J.-H. Kim, "Clinical significance of early detection of esophageal cancer in patients with head and neck cancer," *Gut Liver*, vol. 9, no. 2, pp. 159–165, Mar. 2015.
[5] D. Arnold, W. Goodwin, D. Weed, and F. Civantos, "Treatment of recurrent and advanced stage squamous cell carcinoma of the head and neck," *Seminars Radiat. Oncol.*, vol. 14, no. 2, pp. 190–195, Apr. 2004.
[6] K. Manikantan, S. Khode, R. C. Dwivedi, R. Palav, C. M. Nutting, P. Rhys-Evans, K. J. Harrington, and R. Kazi, "Making sense of post-treatment surveillance in head and neck cancer: When and what of follow-up," *Cancer Treatment Rev.*, vol. 35, no. 8, pp. 744–753, Dec. 2009.
[7] D. G. Deschler and T. Day, "TNM staging of head and neck cancer and neck dissection classification," *Amer. Acad. Otolaryngol. Head Neck Surg. Found.*, pp. 10–23, 2008.
[8] 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, Nov. 2015.
[9] N. S. Murthy and C. Bethala, "Review paper on research direction towards cancer prediction and prognosis using machine learning and deep learning models," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 5, pp. 5595–5613, May 2023.

[10] M. Kazmierski, M. Welch, S. Kim, C. McIntosh, K. Rey-McIntyre, S. H. Huang, T. Patel, T. Tadic, M. Milosevic, F.-F. Liu, A. Ryczkowski, J. Kazmierska, Z. Ye, D. Plana, H. J. W. L. Aerts, B. H. Kann, S. V. Bratman, A. J. Hope, and B. Haibe-Kains, "Multi-institutional prognostic modeling in head and neck cancer: Evaluating impact and generalizability of deep learning and radiomics," *Cancer Res. Commun.*, vol. 3, no. 6, pp. 1140–1151, Jun. 2023.

[11] Y.-C.-A. Lee et al., "Risk prediction models for head and neck cancer in the US population from the INHANCE consortium," *Amer. J. Epidemiol.*, vol. 189, no. 4, pp. 330–342, Apr. 2020.

[12] C. C. Mirestean, O. Pagute, C. G. Buzea, R. I. Iancu, and D. T. Iancu, "Radiomic machine learning and texture analysis—New horizons for head and neck oncology," *Maedica*, vol. 14, no. 2, pp. 126–130, Jun. 2019.

[13] E. R. Gritz, C. R. Carr, D. Rapkin, E. A. Abemayor, L.-J. C. Chang, W. K. Wong, T. Belin, T. Calcaterra, K. T. Robbins, and G. Chonkich, "Predictors of long-term smoking cessation in head and neck cancer patients," *Cancer Epidemiol. Biomarkers Prevention*, vol. 2, no. 3, pp. 261–270, May 1993.

[14] A. Argiris, M. Karamouzis, D. Raben, and R. L. Ferris, "Head and neck cancer," *Lancet*, vol. 371, no. 9625, pp. 1695–1709, May 2008.

[15] D. W. Kim, S. Lee, S. Kwon, W. Nam, I.-H. Cha, and H. J. Kim, "Deep learning-based survival prediction of oral cancer patients," *Sci. Rep.*, vol. 9, no. 1, p. 6994, May 2019.

[16] A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, and J. Seuntjens, "Deep learning in head & neck cancer outcome prediction," *Sci. Rep.*, vol. 9, no. 1, p. 2764, Feb. 2019.

[17] C. Fakhry, A. Psyrri, and A. Chaturvedi, "HPV and head and neck cancers: State-of-the-science," *Oral Oncol.*, vol. 50, no. 5, pp. 353–355, May 2014.

[18] M. L. Welch, S. Kim, A. Hope, S. H. Huang, Z. Lu, J. Marsilla, M. Kazmierski, K. Rey-McIntyre, T. Patel, B. O'Sullivan, J. Waldron, J. Kwan, J. Su, L. S. Ghoraie, H. B. Chan, K. Yip, M. Giuliani, S. Bratman, and T. Tadic. (2023). *Computed Tomography Images from Large Head and Neck Cohort (RADCURE)*. The Cancer Imaging Archive. [Online]. Available: https://doi.org/10.7937/J47W-NM11

[19] G. Ramadevi, K. Naga, R. Usha, and D. Lavanya, "Importance of feature extraction for classification of breast cancer datasets—A study," *Int. J. Sci. Innov. Math. Res.*, vol. 3, no. 2, pp. 763–768, Jul. 2015.

[20] Z. Xu and Z. Wang, "A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and XGBoost ensemble classifier," in *Proc. 11th Int. Conf. Adv. Comput. Intell. (ICACI)*, Jun. 2019, pp. 278–283.

[21] S. Ray, "A quick review of machine learning algorithms," in *Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput. (COMITCon)*, Feb. 2019, pp. 35–39.

[22] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021.

NAMAN DHARIWAL is currently pursuing the B.Tech. degree in computer science engineering from the Vellore Institute of Technology, Vellore. He is an Aspiring Data Science Engineer. His current research interests include artificial intelligence, machine learning, natural language processing, deep learning, and image processing. He is passionate about helping build a just and sustainable world through his knowledge and skill sets.

RITHVIK HARIPRASAD was born in Bengaluru, India, in 2002. He graduated the primary and secondary education from the Delhi Public School, Vasant Kunj, New Delhi, India. He is currently pursuing the B.Tech. degree in computer science and engineering with the Vellore Institute of Technology, Tamil Nadu, India. His current research interests include medical imaging, machine learning, and oncology.

L. MOHANA SUNDARI received the degree in electronics and communication engineering from the Vellore Engineering College, University of Madras, Tamil Nadu, in 2003, the M.E. degree in applied electronics from Anna University, in 2009, and the Ph.D. degree from the Department of Information and Communication, Anna University, Chennai, in 2022. She is currently an Assistant Professor (Senior Grade) with the School of Computer Science and Engineering, Vellore Institute of Technology University, Vellore. She has a teaching experience of more than 17 years. She has published papers in various international journals and conferences. She also published two books on *Antennas* and a book on *Satellite Communication*. Her current research interests include artificial intelligence, image processing, and networking and communication. She is a lifetime member of IEI.