Enhanced Lung and Oesophageal Cancer Detection Using Aggregated Joint Assembly for Decision-making (AJAD): A Novel Classifier

Avijit Kumar Chaudhuri[1], Pramoda Patro[2], Debmalya Mukherjee[3], Ranjan Banerjee[4], Payel Sengupta[5], Amartya Ghosh[6], Soumodip Das[7]

[1]PDF - Computer Science and Engineering, SR University, Warangal, Telangana, India, 506371 & Professor, Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal, India. Email: c.avijit@gmail.com

[2]Director, Centre of AI and Deep Learning, School of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana, India, 506371. Email: pramoda.mtech09@gmail.com

[3]Assistant Professor, Computational Sciences, Brainware University, Barasat, Kolkata – 125, West Bengal, India. Email: dbml.mukherjee@gmail.com
ORCID ID: 0009-0006-9946-0964

[4]Assistant Professor, Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal, India. Email: rbkpccst@gmail.com
ORCID ID: 0009-0003-1950-7530

[5]Assistant Professor, Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal, India. Email: payel9433@gmail.com
ORCID ID: 0000-0003-3981-5971

[6]Assistant Professor, Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal, India. Email: com.amartya@gmail.com
ORCID ID: 0009-0002-2504-3325

[7]Student, CSE AIML 4th year, Techno Engineering College, Banipur. Email: soumo.das2004@gmail.com

Received: 25th May, 2026; **Revised:** 6th June, 2026; **Accepted:** 8th June, 2026; **Available Online:** 09th June, 2026

ABSTRACT

Lung and oesophageal cancer are among the most lethal malignancies in the world. Their treatment faces huge challenges due to high occurrence, poor prognosis, and complicated protocols. Lung cancer is currently the leading cause of cancer-related deaths. However, there have been significant improvements in early diagnosis and precision treatments. AI-driven imaging and machine learning now help determine non-small cell lung cancer (NSCLC) subtypes without extra staining. The rise in lung cancer among never smokers, potentially due to factors like air pollution, points to environmental risks beyond tobacco. New targeted therapies, such as HER2 and c-Met inhibitors, offer extra precision oncology options for specific NSCLC subgroups.

Oesophageal cancer, mainly oesophageal squamous cell carcinoma (ESCC) and oesophageal adenocarcinoma (EAC), remains an important global health burden. It tends to be diagnosed at an advanced stage, resulting in poor survival rates despite multimodal intervention. Most studies on oesophageal cancer focus on optimizing existing therapies, including minimally invasive surgery and post-treatment quality of life interventions. Additionally, high rates of comorbidity with pulmonary complications and the presence of synchronous tumors in the lung and esophagus highlight the need for integrated oncological therapy. Progress in creating predictive nomograms and guidelines to assess prognosis in such cases is a significant advancement toward better patient outcomes.

Recent statistics show that lung and oesophageal cancers share risk factors and overlap clinically. This highlights the need for integrated research. Oxidative stress, chronic inflammation, and immune disregard, as seen in diseases like COPD, play a central role in lung carcinogenesis. These factors may also alter the tumor microenvironment of both thoracic organs. Rare cases of dual primary malignancies, such as oesophageal malignant melanoma with lung adenocarcinoma, show the complexity of treating co-existing tumors.

Nevertheless, the connections between the pathogenesis of lung and oesophageal cancer, shared risk factors in the environments and genetics, and the difficulties of their concomitant treatment still have gaps. There is an urgent need to integrate screening processes, expand molecular profiling across both tumor types, and develop effective therapeutic options to address these complexities. To address this, this paper introduces an Aggregated Joint Assembly of Decision-making (AJAD) classifier that combines Random Forests, Gradient Boosting, and Extra Trees via soft voting to improve diagnostic accuracy and support more robust clinical decision-making for both cancers. By leveraging this classifier, clinicians may achieve more accurate patient-specific risk stratification and treatment planning through integration of bioinformatics, oncology, and thoracic surgery expertise. AI-guided diagnostics and biomarker-based therapies are expected to be central in advancing clinical paradigms and improving survival rates and quality of life in patients with lung and oesophageal cancers.

Keywords: lung cancer, oesophageal cancer, AI diagnostics, precision medicine, NSCLC, ESCC, comorbidity, molecular profiling, prognosis.

How to cite this article: Chaudhuri AK, Patro P, Mukherjee D, Banerjee R, Sengupta P, Ghosh A, Das S. Enhanced Lung and Oesophageal Cancer Detection Using Aggregated Joint Assembly for Decision-making (AJAD): A Novel Classifier. Int J Drug Deliv Technol. 2026;16(57s): 1596--1615. DOI: 10.25258/ijddt.16.57s.159
Source of support: Nil.
Conflict of interest: None.

1. Introduction

Lung and oesophageal cancer are among the world's most fatal cancers. Combined, they account for millions of cases and cancer-related deaths every year. Lung cancer is the most common cause of cancer death worldwide. It has an annual incidence of 2-2.5 million new cases and a five-year survival rate of only 5 percent, due to late diagnosis (Smolarz et al. 2025). The growth of molecular oncology has begun to improve diagnostic and treatment approaches, especially for non-small cell lung cancer (NSCLC). However, problems remain with early diagnosis and treatment resistance. Tobacco smoking continues to be a major risk factor. Recent findings also implicate environmental exposures, particularly air pollution, in lung carcinogenesis among never smokers. Studies have found strong links between particulate pollution and DNA mutations, such as TP53. This suggests that non-tobacco factors are becoming increasingly important globally (Zhong et al., 2025; Hua et al., 2025).

On the other hand, oesophageal cancer is divided into two major histological types: oesophageal squamous cell carcinoma (ESCC) and oesophageal adenocarcinoma (EAC). Both types have poor prognoses, mainly because they progress without symptoms and are often diagnosed late. ESCC is more prevalent in Asian and African regions with high incidence, whereas EAC has risen in the Western population following the growing prevalence of gastroesophageal reflux disease (GERD) and obesity. Five-year survival is still poor even with recent advancements in surgical procedures and chemoradiation regimens. The current global studies bring into light the fact that oesophageal cancer remains a significant burden to health because of high mortality rates and little advancements in long-term survival gains(Jiang et al., 2025; Langley et al., 2025; Chaudhuri et al., 2018).

Lung cancer is a heterogeneous disease with many molecular subtypes that affect prognosis and treatment. Most cases are NSCLC. This category includes adenocarcinoma, squamous cell carcinoma, and large cell carcinoma, each with specific genetic changes. Targeted therapies for mutations in EGFR, ALK, and HER2 have transformed the management of subtypes. Zongertinib, a HER2 tyrosine kinase inhibitor, and telisotuzumab vedotin, a c-Met-directed antibody-drug conjugate, are now approved for some NSCLC populations. These therapies expand precision oncology options (Langley et al., 2025; Das et al., 2025).

Oesophageal cancer is less responsive to molecularly targeted therapy. Current research is clarifying actionable pathways and immunotherapy combinations. This cancer shows genomic instability, immune escape, and aggressive local invasion. As a result, treatment requires multimodal surgery and chemoradiation. Prehabilitation interventions focus on optimizing patients' health before extensive thoracic surgery in both lung and oesophageal cancer cases. These interventions reduce complications and improve survival (Chaudhuri et al., 2020; Chaudhuri et al., 2023; Das et al., 2026).

Lung and oesophageal cancers are different diseases, yet they have overlapping risk factors and clinical features. Tobacco smoking is the main shared risk factor for both ESCC and lung cancer. Chronic inflammation, whether from COPD or GERD, promotes disease progression. COPD-lung cancer comorbidity research shows that persistent inflammation creates microenvironments that promote tumors. This highlights the importance of common pathways.

Patients with one primary tumor in the thoracic organs have a higher risk of developing a second primary tumor nearby. Systematic reviews indicate that oesophageal cancer patients are also at risk for a second primary lung tumor, and vice versa, though the prevalence is low. These overlaps highlight the need for detailed risk stratification and surveillance in high-risk populations (van Tilburg et al., 2023).

Early diagnosis is a challenge for both cancers. AI-aided imaging and predictive models can improve diagnostic accuracy and guide therapy choices in lung cancer, especially in complex NSCLC cases. Machine learning models that combine imaging, genomic, and clinical data help predict therapy resistance and support personalized treatments. Diagnostics for oesophageal cancer have improved with the development of endoscopic imaging and biomarkers, but survival rates have not increased significantly. Surgical innovations, like minimally invasive esophagectomy, improve outcomes. Research now also focuses on patient-centered measures and classic endpoints such as survival (Yang & Yang, 2025; Bidzińska & Szurowska, 2023; Diaz-Gay et al., 2025).

Despite recent progress, gaps remain in knowledge about factors underlying lung and oesophageal cancers, both shared and independent. Future research should focus on: • Combined screening processes for high-risk groups. • Broader molecular characterization of esophageal tumors to expand targeted therapy options.

• Develop prognostic models for complex clinical cases with either synchronous or metastatic disease.
• Expand investigation of occupational and non-tobacco risks from the environment and lifestyle. Such insights should be translated into improved clinical outcomes through an interdisciplinary framework that integrates oncology, surgery, computational biology, and public health.

The focus of treatment strategies tailored to each patient's molecular, genetic, and tumor features is a key element of comprehensive cancer therapy, and such a solution is an excellent fit for AI-based approaches. Oncology has seen significant interest in the use of AI, specifically machine learning (ML), to support activities such as risk assessment, diagnosis, drug discovery, and molecular profiling of tumors. It has been established that the use of ML improves the predictive and diagnostic capabilities of cancer analysis compared with traditional cognitive-based analysis of pathology micrographs and imaging studies, which tend to transform images into numeric sequences (Das et al., 2025).

1. Relevant Literature

Table 1 summarizes accuracy-related literature research and comparative statistical analysis for lung cancer.

Table 1. Comparison of Performance with Previous Studies

Reference	**Problem Type**	**Dataset / Features Used**	**Machine Learning Technique**	**Accuracy (%)**	**Other Reported Results**
Singh (2024)	Lung cancer onset prediction	Demographic + medical history features	LR	87–88	AUC-ROC reported, Sensitivity & F1 discussed
			SVM	~90–91	Improved sensitivity over LR
			RF	~92	Best overall performance among ML models
Ghosh & Bhattacharjee, (2024)	Lung cancer diagnosis	Kaggle dataset (3001 samples, 16 features)	KNN	~94	Precision and Recall reported
			DT	~95	Balanced performance
			SVM	~96	AUC-ROC ≈ 0.95
			RF	~97–98	Highest accuracy, best F1-score
Farshchiha et al., 2025	Lung cancer stage/level classification	Clinical + CT-derived features	Logistic Regression	~97–98	High recall and precision
			Random Forest	~99	Stable across folds
			XGBoost	≈99–100	Best ML performance
			LightGBM	≈99–100	Comparable to XGBoost
			Deep Learning (Transfer Learning CNNs)	~92–94	Underperformed ML due to limited data
Abdullah et al., 2025	Lung tumor detection	Public lung cancer dataset	CNN (baseline)	~97–98	Precision & Recall reported
			FFXO + Bi-GAN (Proposed)	98.7	Improved F1-score and robustness

Chan (2024)	Lung cancer diagnosis	Kaggle risk-factor dataset	Neural Network	50–60	Diagnosis prediction unreliable
		Lung cancer severity prediction	Severity dataset	Neural Network	~95–99
Present Work	**Lung cancer diagnosis**	**All available clinical features (no feature reduction)**	**LR, RF, ET, GBD, AdaBoost, Voting Ensemble**	**96**	**Sensitivity = 96, Specificity = 97, Precision = 97, F1 = 96, Kappa = 0.71, AUC = 0.89**

2. Assessment of Model Performance

Dataset Description

Table 3. Descriptive Statistics of Variables (Lung Cancer Dataset)

Column Name	Data Type	Null Values	Non-Null Values	Min	Max	Mean	Std Dev	25th Percentile	50th Percentile	75th Percentile
AGE	int64	0	309	21	87	62.67	8.21	57	62	69
SMOKING	int64	0	309	1	2	1.56	0.5	1	2	2
YELLOW_FINGERS	int64	0	309	1	2	1.57	0.5	1	2	2
ANXIETY	int64	0	309	1	2	1.5	0.5	1	2	2
PEER_PRESSURE	int64	0	309	1	2	1.5	0.5	1	2	2
CHRONIC DISEASE	int64	0	309	1	2	1.5	0.5	1	2	2
FATIGUE	int64	0	309	1	2	1.67	0.47	1	2	2
ALLERGY	int64	0	309	1	2	1.56	0.5	1	2	2
WHEEZING	int64	0	309	1	2	1.56	0.5	1	2	2
ALCOHOL CONSUMING	int64	0	309	1	2	1.56	0.5	1	2	2
COUGHING	int64	0	309	1	2	1.58	0.49	1	2	2
SHORTNESS OF BREATH	int64	0	309	1	2	1.64	0.48	1	2	2
SWALLOWING DIFFICULTY	int64	0	309	1	2	1.47	0.5	1	1	2
CHEST PAIN	int64	0	309	1	2	1.56	0.5	1	2	2
LUNG_CANCER	int64	0	309	0	1	0.87	0.33	1	1	1

The Lung Cancer Dataset used in this study is a comprehensive and well-organized resource

(https://www.kaggle.com/datasets/akashnath29/lung-cancer-dataset). It serves as a vital asset for developing sophisticated systems for cancer detection, especially in AI-based health technologies, and is summarized in Table 3. The dataset involves a wide range of symptoms, patient demographics, and clinical data. These details are critical for building powerful models that precisely predict and diagnose lung cancer. The dataset is holistic because it contains the most important information, ensuring the final model is versatile and accurate.

The records begin with patient demographics, including age, gender, and smoking status. These are basic variables in investigating the correlation with lung cancer. Age allows the examination of age-related incidences. Gender enables easy gender-based studies. Smoking status—whether the patient is a current smoker, former smoker, or non-smoker—is important for determining the effects of smoking on lung cancer risk and progression.

Another valuable part of the dataset is the medical history. Particular attention is paid to comorbidities such as chronic obstructive pulmonary disease (COPD), as this information is highly useful for treatment planning and prognosis. This supplementary health information develops a more detailed picture of factors that can affect lung cancer patient outcomes.

Clinical data include vital signs such as blood pressure, heart rate, and respiratory rate. These measurements provide a clear picture of the patient's health during diagnosis and throughout treatment. They are essential for continuous assessment of health status and for making informed clinical judgments.

The datasets also selectively include a significant range of symptoms common to lung cancer. These include fatigue, coughing, wheezing, shortness of breath, chest pain, and difficulty swallowing. There are also lifestyle factors such as yellow fingers, anxiety, peer pressure, chronic diseases, allergy, and alcohol consumption. This variety provides a broad picture of symptoms that can help with early detection.

This dataset requires several important pre-processing steps: cleaning, normalization, and tokenization. Data cleaning ensures that no irrelevant or redundant entries are left behind, whereas normalization makes text data uniform by converting it to lowercase and removing unnecessary characters. The symptoms and conclusion are separated into tokens, or individual words, to be easier to analyze and feed into machine learning models.

This data applies across a broad spectrum in the medical and educational spheres. It enables the creation of machine learning models to aid in early disease detection, predict disease progression, and provide individualized therapy plans. It can also be analyzed statistically to reveal patterns and correlations in the development of lung cancer. Furthermore, the data can be used by healthcare providers to improve patient care and optimize treatment regimens, while researchers can use it to conduct in-depth research on the pathophysiology of lung cancer, its risk factors, and its response to treatment.

The Lung Cancer Dataset is accessible in multiple formats, including JSON and CSV. Its compatibility makes it useful with many data science and healthcare systems. The dataset's high content and variety make it an effective resource for research on lung cancer and for enhancing AI-powered health technologies.

To sum up, the Lung Cancer Dataset is a valuable resource for researchers and medical practitioners in the fight against lung cancer. It contains extensive patient demographics, clinical information, symptoms, and medical history. This strong foundation helps develop accurate, AI-based predictive models. Using this dataset, the healthcare community can make significant progress in improving diagnosis and patient outcomes.

Table 4. Descriptive Statistics of Variables (Oesophageal Cancer)

Column Name	Data Type	Null	Non-Null Values	Min	Max	Mean	Std Dev
patient_barcode	object	0	3985	-	-	-	-
tissue_source_site	object	0	3985	-	-	-	-
patient_id	object	0	3985	-	-	-	-
bcr_patient_uuid	object	0	3985	-	-	-	-
informed_consent_verified	object	0	3985	-	-	-	-
icd_o_3_site	object	0	3985	-	-	-	-
icd_o_3_histology	object	0	3985	-	-	-	-
icd_10	object	0	3985	-	-	-	-
tissue_prospective_collection_indicator	object	40	3945	-	-	-	-
tissue_retrospective_collection_indicator	object	40	3945	-	-	-	-
days_to_birth	int64	0	3985				

country_of_birth	object	1927	2058	-	-	-	-
gender	object	0	3985	-	-	-	-
height	float64	219	3766	145	202	172.13	9.08
weight	float64	40	3945	41	198	75.62	19
country_of_procurement	object	40	3945	-	-	-	-
state_province_of_procurement	object	1280	2705	-	-	-	-
city_of_procurement	object	860	3125	-	-	-	-
race_list	object	419	3566	-	-	-	-
ethnicity	object	2048	1937	-	-	-	-
other_dx	object	0	3985	-	-	-	-
history_of_neoadjuvant_treatment	object	0	3985	-	-	-	-
person_neoplasm_cancer_status	object	335	3650	-	-	-	-
vital_status	object	0	3985	-	-	-	-
days_to_last_followup	float64	1197	2788	-4	374	306.2	506.18
days_to_death	float64	278	1197	0	214	467.86	456.2
tobacco_smoking_history	float64	380	3605	1	4	2.36	1.14
age_began_smoking_in_years	float64	2253	1732	10	57	22.21	7.61
stopped_smoking_year	float64	2377	1608	1949	2013	1987.81	19.21
number_pack_years_smoked	float64	1816	2169	1	102	35.39	21.61
alcohol_history_documented	object	60	3925	-	-	-	-
frequency_of_alcohol_consumption	float64	1516	2469	0	7	3.52	3.13
amount_of_alcohol_consumption_per_day	float64	1877	2108	0	14	1.75	2.23
reflux_history	object	677	3308	-	-	-	-
antireflux_treatment_types	object	2988	997	-	-	-	-
h_pylori_infection	object	2428	1557	-	-	-	-
initial_diagnosis_by	object	738	3247	-	-	-	-
barretts_esophagus	object	818	3167	-	-	-	-
goblet_cells_present	object	3566	419	-	-	-	-
history_of_esophageal_cancer	object	837	3148	-	-	-	-
number_of_relatives_diagnosed	float64	3146	839	0	2	0.33	0.56
has_new_tumor_events_information	object	0	3985	-	-	-	-
day_of_form_completion	int64	0	3985	1	30	16.47	8.12

month_of_form_completion	int64	0	3985	1	12	4.81	3.74
year_of_form_completion	int64	0	3985	2012	2015	2013.55	0.6
has_follow_ups_information	object	0	3985	-	-	-	-
has_drugs_information	object	0	3985	-	-	-	-
has_radiations_information	object	0	3985	-	-	-	-
project	object	0	3985	-	-	-	-
stage_event_system_version	object	0	3985	-	-	-	-
stage_event_clinical_stage	object	2667	1318	-	-	-	-
stage_event_pathologic_stage	object	498	3487	-	-	-	-
stage_event_tnm_categories	object	0	3985	-	-	-	-
stage_event_psa	float64	3985	0				
stage_event_gleason_grading	float64	3985	0				
stage_event_ann_arbor	float64	3985	0				
stage_event_serum_markers	float64	3985	0				
stage_event_igcccg_stage	float64	3985	0				
stage_event_masaoka_stage	float64	3985	0				
primary_pathology_tumor_tissue_site	object	0	3985	-	-	-	-
primary_pathology_esophageal_tumor_cental_location	object	20	3965	-	-	-	-
primary_pathology_esophageal_tumor_involvement_sites	object	20	3965	-	-	-	-
primary_pathology_histological_type	object	0	3985	-	-	-	-
primary_pathology_columnar_metaplasia_present	object	1595	2390	-	-	-	-
primary_pathology_columnar_mucosa_goblet_cell_present	object	2174	1811	-	-	-	-
primary_pathology_columnar_mucosa_dysplasia	object	2235	1750	-	-	-	-
primary_pathology_neoplasm_histologic_grade	object	0	3985	-	-	-	-
primary_pathology_days_to_initial_pathologic_diagnosis	int64	0	3985	0	0	0	0
primary_pathology_age_at_initial_pathologic_diagnosis	int64	0	3985	27	90	63.48	12.18
primary_pathology_year_of_initial_pathologic_diagnosis	float64	140	3845	1998	2013	2009.24	4.2
primary_pathology_initial_pathologic_diagnosis_method	object	100	3885	-	-	-	-
primary_pathology_init_pathology	object	3106	879	-	-	-	-

_dx_method_other							
primary_pathology_lymph_node_metastasis_radiographic_evidence	object	837	3148	-	-	-	-
primary_pathology_primary_lymph_node_presentation_assessment	object	3200	3665	-	-	-	-
primary_pathology_lymph_node_examined_count	float64	10000	2985	1	87	14.27	12.19
primary_pathology_number_of_lymphnodes_positive_by_he	float64	10000	2985	0	21	2.45	3.32
primary_pathology_number_of_lymphnodes_positive_by_ihc	float64	2533	1452	0	9	0.29	1.25
primary_pathology_planned_surgery_status	object	2507	1478	-	-	-	-
primary_pathology_treatment_prior_to_surgery	object	28447	1138	-	-	-	-
primary_pathology_residual_tumor	object	5200	3465	-	-	-	-
primary_pathology_karnofsky_performance_score	float64	2625	1360	20	100	73.82	16.1
primary_pathology_eastern_cancer_oncology_group	float64	2628	1357	0	3	1.02	0.74
primary_pathology_radiation_therapy	object	6338	3347	-	-	-	-
primary_pathology_postoperative_rx_tx	object	6583	3327	-	-	-	-

The Oesophageal Cancer Dataset (https://www.kaggle.com/datasets/abhinaba1biswas/esophageal-cancer-dataset) used in this research is a valuable resource for medical care providers and scientists focused on cancer diagnosis, individualized therapies, and prognosis algorithms. It provides an extensive set of real-world clinical data—a strong baseline for AI-based solutions to enhance treatment and diagnosis. This comprehensive dataset includes diverse patient and tumor features, treatment history, and patient outcomes, all crucial to understanding oesophageal cancer progression. Additional information can be viewed in Table 4.

The data begins with Patient Demographics, which provides essential details about the patient's identity and background. It comes with the Patient Barcode, a unique identifier for each patient that helps ensure the privacy and integrity of their data. The Tissue Source Site is a code indicating where the tissue sample was obtained, providing background for the tumor analysis. The Age at Diagnosis feature is essential for conducting age-based research to investigate the relationship between age and cancer incidence or outcomes. Additionally, the gender attribute will enable gender-specific analysis to reveal potential differences in disease progression between female and male patients. Informed Consent Verified field: Indicates whether the patient has given informed consent, ensuring that ethical standards are observed when collecting data and conducting research.

The Medical and Clinical History section of the dataset contains essential data on the patient's background, providing relevant insights for clinicians. ICD-10 and ICD-O-3 Codes present standardized information on cancer site and histology (e.g., squamous cell carcinoma or adenocarcinoma) to support an accurate understanding of tumor characteristics. Documentation of co-morbidities, such as Gastroesophageal Reflux Disease (GERD), is included, since these can influence treatment outcomes and prognosis. Smoking Status—recorded as current, former, or non-smoker—is captured to assess how tobacco use impacts the risk and progression of oesophageal cancer.

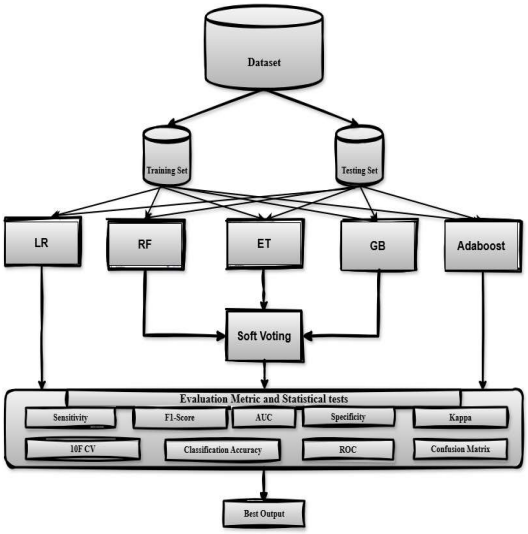
For Cancer-Specific Data, the dataset offers detailed insights into tumor characteristics and progression. Tumor Location specifies whether the tumor is upper, middle, or lower, aiding treatment planning and prognosis. Histology categorizes cancer type (such as squamous cell carcinoma or adenocarcinoma), supporting tailored treatment approaches. Cancer Stage reflects the stage at

diagnosis (Stage 0-IV), impacting treatment decisions and prognosis. Residual Tumor Status indicates whether residual tumor remains after surgery, classified as R0 (none) or R1 (present), and informs surgical success and the need for additional intervention. Lymph Node Examination shows how many nodes were examined and how many were metastatic, clarifying the extent of cancer spread. Radiation Therapy and Postoperative Treatment fields document if the patient is receiving radiation or other postoperative care, helping track treatment regimens and outcomes.

The Clinical Outcome Data in the dataset evaluate the patient's health and daily activity. The Karnofsky Performance Score measures overall functional status and activity. The Eastern Cooperative Oncology Group (ECOG) Performance Status gives more insight into the patient's ability to manage cancer and treatment.

Such a rich dataset will provide a solid foundation for AI-based cancer detection and prognostic tools. It supports machine learning applications that predict patient outcomes, customize treatment strategies, and improve clinical decision-making. By incorporating extensive patient demographics, tumor and treatment histories, and clinical outcomes, the dataset helps healthcare providers and researchers better manage oesophageal cancer, advancing research and treatment. The proposed cancer prediction framework is depicted in Figure 1 and illustrated in Algorithm 1 below.

Figure 1. Cancer Prediction Framework



Algorithm 1: Soft Voting with Specific Models

Step 1: Notation and Definitions

Dataset

Let

$$D = \{(x_n, y_n)\}_{n=1}^{N}$$

where

- $x_n \in R^d$ is the **feature vector**
- $y_n \in \{0,1\}$ is the **binary class label**

Performance Parameters

Let

- **Accu ∈ [0,1]** = accuracy
- **Spec ∈ [0,1]** = specificity
- **Sens ∈ [0,1]** = sensitivity

These parameters are used to **tune model decision thresholds or loss weighting**.

Step 2: Model Initialization (Iteration 1)

For iteration index

$$t = 1, \ldots, 10$$

define four classifiers:

$$R_t = Random(Spec, Sens, Accu)$$
$$E_t = ExtraTree(Spec, Sens, Accu)$$
$$G_t = GB(Spec, Sens, Accu)$$
$$L_t = LR(Spec, Sens, Accu)$$
$$A_t = AdaBoost(Spec, Sens, Accu)$$

Let the complete pool of trained models be:

$$M = \{M_1, M_2, \ldots, M_{10}\}$$

where each M_i is a classifier producing a posterior probability:

$$M_i(x) = P(y = 1 \mid x, M_i)$$

Step 3: Model Combination Search (Iterations 2–4)

Define index sets:

$$i, j, k \in \{1,2,3,4,5\}$$

subject to constraints:

$$i \neq j, i \neq k, j \neq k$$
$$i \leq j \leq k$$

Step 4: Soft Voting Function

Given three distinct models M_a, M_b, M_c, define the Soft Voting output as:

$$\hat{p}(x \mid M_a, M_b, M_c) = \frac{1}{3}(M_a(x) + M_b(x) + M_c(x))$$

Final predicted class

$$\hat{y}(x) = \begin{cases} 1 & \text{if } \hat{p}(x) \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

where

τ= threshold chosen based on specificity–sensitivity trade-off.

Step 5: Soft Voting Evaluation Metric

Let the performance metric of a soft voting ensemble be:

$$SV(M_a, M_b, M_c) = F(Sens, Spec)$$

A common formulation is:

$$F = \alpha \cdot Sens + (1 - \alpha) \cdot Spec$$

where

$$\alpha \in [0,1]$$

Step 6: Initialization of Best Score

Initialize:

$$SV^* = -1$$

This variable stores the maximum soft voting score found so far.
Soft Voting Computation
For each valid triple (i, j, k):

$$SV_{i,j,k} = SV(M_i, M_j, M_k)$$

Update Rule

$$SV^* = \max(SV^*, SV_{i,j,k})$$

Step 7: Complete Optimization Objective
The algorithm solves:

$$SV^* = \max_{i,j,k \in \{1,\ldots,5\}} SV(M_i, M_j, M_k)$$

subject to

$$i \neq j \neq k$$
$$i \leq j \leq k$$

Step 8: Final Outcome

- All valid 3-model combinations are exhaustively evaluated
- Soft voting scores are computed for each combination
- The maximum soft voting score is stored in SV^*

Thus,
SV^* presents the best-performing ensemble under the given specificity-sensitivity constraints.

Explanation for Higher Accuracy:

The diagram above (Figure 1) depicts an ensemble learning framework using several machine learning classifiers: Logistic Regression (LR), Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB), and AdaBoost. These classifiers are then combined using a Soft Voting technique, which averages their predicted probabilities, to produce a final output known as the classifier Aggregated Joint Assembly for Decision-making (AJAD). This integrated process yields the highest accuracy among all the techniques.

To clarify why Soft Voting achieves the highest accuracy, we will first break down the technique in greater detail. Afterward, we will discuss its mathematical rationale and explain why it is advantageous in this situation.

Before delving deeper, recall that Soft Voting in ensemble learning determines the final prediction by combining the predicted probabilities of each class label from individual classifiers, rather than choosing the class with the most votes, as in hard voting. The key idea is that:

- Each classifier generates a probability distribution, which is a list of probabilities that sum to 1, over the set of possible classes.
- The average (mean) of these probability distributions is then computed for each class across all the classifiers.
- The class with the highest combined average probability is selected as the final outcome.

Mathematically, soft voting can be defined as:

$$P(y = c \mid X) = \frac{1}{N} \sum_{i=1}^{N} P_i(y = c \mid X)$$

Where:

- P(y=c|X)P(y = c | X)P(y=c|X) is the final probability for class c given input X.
- Pi(y=c|X) is the probability predicted by classifier i for class c given input X.
- N is the total number of classifiers in the ensemble.

The class C with the highest probability will be selected as the final prediction.

1. Diversity of Classifiers: Various classifiers are trained on data in different ways, and each has its pros and cons. For example:

- Logistic regression can use a linear decision boundary.
- Random Forest and Extra Trees can work well where the data is non-linear.
- Gradient Boosting and AdaBoost are likely to work with weak learners and decrease bias.
- AJAD achieves this by taking the union of the predictions made by several classifiers, but each of the classifiers has its strengths and weaknesses, which the other one can have a chance to overcome. Overall predictive accuracy results from combining multiple classifiers.
- With soft voting, particularly on high-performing classifiers such as Gradient Boosting and AdaBoost, improved overall generalization occurs because the ensemble averages out each classifier's errors. This alleviates overfitting, which can be larger in models trained on particular subsets of the data or be very sensitive to noise.

The Central Limit Theorem: Another way to explain the performance improvement is through the Central Limit Theorem (CLT), which states that, when collecting independent estimators (including classifiers, in this case), the error will be smaller (i.e., the predictions will be more stable and reliable). Since individual classifiers have their own bias and variance, combining them can help counter such errors.

Combining Probabilities: Soft voting combines probabilities rather than producing final choices (labels). This enables the model to take into account not only the majority's opinion, but also each classifier's degree of confidence. An example of this is when one classifier is very sure yet wrong; soft voting can still under-weight its vote and over-weight classifiers who may be less sure but more accurate.

Mathematically, this is the weighted

average of the classifiers' outputs. If the classifiers are well-calibrated, their accuracies and reliabilities will be combined to produce the final prediction.

- Classifier Combination: AJAD likely refers to the final classifier obtained via soft voting, combining RF, ET, and GB. Suppose that every classifier is such that its probability distribution is different in each class. These probabilities are averaged to yield the final probability for each class. The classifiers' biases (or errors) do not show perfect correlation and differ across hypotheses. The result of the ensemble will be stronger and less prone to overfitting or underfitting due to:
- Each classifier has a different error pattern and sensitivity to the data.
- Soft voting aggregates the predictions of multiple classifiers to improve overall performance.

1. Model Calibration: It is important to calibrate the classifiers (i.e., ensure their prediction probabilities are accurate) in soft voting. When the individual classifiers are well calibrated (e.g., via Platt scaling or isotonic regression), averaging their probabilities yields a better estimate of the true class probabilities.
2. Statistical Performance Metrics to Use: The diagram also includes various metrics for evaluating classifier performance, such as Sensitivity, F1-Score, AUC, Classification Accuracy, and ROC Curve. These measures are used to make sure that the compound classifier AJAD is tested not only on its ability to be accurate but also on its ability to generalize to various areas of performance, such as:
   Sensitivity (True Positive Rate): It is a measure of how well a classifier detects positive instances.
   F1-Score: This metric combines precision (the proportion of true positives among predicted positives) and recall (the proportion of true positives detected), making it effective with imbalanced datasets.
3. AUC (Area Under the Curve): This metric measures the model's ability to differentiate between classes.
4. Classification Accuracy: The percentage of accurate predictions made by the classifier. Considering that AJAD relies on soft voting, it will likely perform well across various metrics by leveraging the strengths of other classifiers, which are more resistant to data changes and tend to avoid underfitting or overfitting.
5. Lastly, the AJAD classifier obtained using the Soft Voting technique is a fusion of the strengths of RF, ET, and GB. The high accuracy is due to:
6. The variation among the classifiers within the ensemble and among individuals yields different interpretations of the data.
7. Probability-based decision-making (soft voting) instead of simple majority voting: It can minimize bias and variance by combining predictions from several models.
   The AJAD classifier can make the most accurate prediction, particularly in terms of the metrics it uses to gauge the prediction (e.g., AUC, F1-Score, ROC).

3. Results and Discussion

Table 5. Model Accuracy Across Folds (Lung Cancer Dataset)

Fold	Model	Accuracy	Sensitivity	Specificity	Precision	F1	Kappa	AUC
1	RandomForest	0.97	1.00	0.75	0.96	0.98	0.84	1.000
1	GradientBoosting	0.90	1.00	0.25	0.90	0.95	0.37	0.98
1	ExtraTrees	0.97	1.00	0.75	0.96	0.98	0.84	1.000
1	AdaBoost	0.94	1.00	0.50	0.93	0.96	0.64	0.98
1	LogisticRegression	0.87	1.00	0.00	0.87	0.93	0.000	0.81
1	AJAD	0.97	1.00	0.75	0.96	0.98	0.84	1.000
2	RandomForest	0.90	0.96	0.50	0.93	0.95	0.52	0.81
2	GradientBoosting	0.94	1.00	0.50	0.93	0.96	0.64	0.84

2	Extra Trees	0.87	0.93	0.50	0.93	0.93	0.43	0.80
2	AdaBoost	0.87	0.93	0.50	0.93	0.93	0.43	0.90
2	LogisticRegression	0.87	0.96	0.25	0.90	0.93	0.27	0.88
2	AJAD	0.94	1.00	0.50	0.93	0.96	0.64	0.87
3	RandomForest	0.87	0.93	0.50	0.93	0.93	0.43	0.94
3	GradientBoosting	0.94	0.96	0.75	0.96	0.96	0.71	0.94
3	Extra Trees	0.84	0.89	0.50	0.92	0.91	0.35	0.94
3	AdaBoost	0.90	0.89	1.00	1.00	0.94	0.67	0.95
3	LogisticRegression	0.90	1.00	0.25	0.90	0.95	0.37	0.96
3	AJAD	0.90	0.93	0.75	0.96	0.94	0.61	0.95
4	RandomForest	0.94	1.00	0.50	0.93	0.96	0.64	0.87
4	GradientBoosting	0.90	1.00	0.25	0.90	0.95	0.37	0.80
4	Extra Trees	0.90	1.00	0.25	0.90	0.95	0.37	0.83
4	AdaBoost	0.94	1.00	0.50	0.93	0.96	0.64	0.93
4	LogisticRegression	0.94	1.00	0.50	0.93	0.96	0.64	0.88
4	AJAD	0.90	1.00	0.25	0.90	0.95	0.43	0.86
5	RandomForest	0.97	1.00	0.75	0.96	0.98	0.84	1.00
5	GradientBoosting	0.97	1.00	0.75	0.96	0.98	0.84	0.99
5	Extra Trees	0.97	1.00	0.75	0.96	0.98	0.84	1.00
5	AdaBoost	0.90	0.96	0.50	0.93	0.95	0.52	0.95
5	LogisticRegression	0.90	1.00	0.25	0.90	0.95	0.37	0.88
5	AJAD	0.97	1.00	0.75	0.96	0.98	0.84	1.00
6	RandomForest	0.87	0.93	0.50	0.93	0.93	0.43	0.87
6	GradientBoosting	0.81	0.89	0.25	0.89	0.89	0.14	0.81
6	Extra Trees	0.81	0.85	0.50	0.92	0.88	0.29	0.78
6	AdaBoost	0.74	0.81	0.25	0.88	0.85	0.05	0.75
6	LogisticRegression	0.84	0.93	0.25	0.89	0.91	0.20	0.65
6	AJAD	0.84	0.89	0.50	0.92	0.91	0.35	0.84
7	RandomForest	0.84	0.89	0.50	0.92	0.91	0.35	0.88

7	GradientBoosting	0.84	0.89	0.50	0.92	0.91	0.35	0.82
7	ExtraTrees	0.90	0.96	0.50	0.93	0.95	0.52	0.77
7	AdaBoost	0.74	0.74	0.75	0.95	0.83	0.30	0.84
7	LogisticRegression	0.77	0.89	0.00	0.86	0.87	-0.12	0.69
7	AJAD	0.84	0.89	0.50	0.92	0.91	0.35	0.84
8	RandomForest	0.90	0.93	0.75	0.96	0.94	0.61	0.93
8	GradientBoosting	0.87	0.93	0.50	0.93	0.93	0.43	0.94
8	ExtraTrees	0.94	0.93	1.00	1.00	0.96	0.76	0.94
8	AdaBoost	0.90	0.93	0.75	0.96	0.94	0.61	0.94
8	LogisticRegression	0.87	0.96	0.25	0.90	0.93	0.27	0.72
8	AJAD	0.94	0.93	1.00	1.00	0.96	0.76	0.93
9	RandomForest	0.97	1.00	0.75	0.96	0.98	0.84	0.98
9	GradientBoosting	0.97	1.00	0.75	0.96	0.98	0.84	0.96
9	ExtraTrees	0.97	1.00	0.75	0.96	0.98	0.84	0.99
9	AdaBoost	0.97	0.96	1.00	1.00	0.98	0.87	0.99
9	LogisticRegression	0.90	1.00	0.25	0.90	0.95	0.37	0.96
9	AJAD	0.97	1.00	0.75	0.96	0.98	0.84	0.98
10	RandomForest	0.93	0.96	0.67	0.96	0.96	0.63	0.98
10	GradientBoosting	0.93	0.93	1.00	1.00	0.96	0.71	0.99
10	ExtraTrees	0.97	0.96	1.00	1.00	0.98	0.84	0.97
10	AdaBoost	0.97	0.96	1.00	1.00	0.98	0.84	1.00
10	LogisticRegression	0.93	1.00	0.33	0.93	0.96	0.47	0.99
10	AJAD	0.97	0.96	1.00	1.00	0.98	0.84	0.99

The research paper compares five machine learning models—Random Forest, Gradient Boosting, Extra Trees, AdaBoost, and Logistic Regression—with the AJAD model across 10-fold cross-validation on a lung cancer dataset. Key evaluation metrics—Accuracy, Sensitivity, Specificity, Precision, F1 Score, Kappa, and AUC—capture predictive performance. The discussion presents each model's ability to detect lung cancer, with classification results summarized in Table 5.

Table 5: Confusion matrix, performance evaluation metrics, and statistical tests

S/N	Metrics	Formula/ Description			
1	Confusion Matrix			Actual	
				Malignant (Positive)	Benign (Negative)

			Malignant (Positive)	True Positive, T_P	False Positive, F_P
			Benign (Negative)	False Negative, F_N	True Negative, T_N
				Sensitivity=	Specificity=
2	Accuracy	$\frac{T_P + T_N}{(T_P + F_N) + (F_P + T_N)}$			
3	Precision	$\frac{T_P}{T_P + F_P}$			
4	AUC (Area under the curve)	A curve plotted between sensitivity and (1-specificity) is called receiver operating characteristic (ROC). AUC measures the degree to which the curve is up in the north-west corner.			
5	Kappa Statistic	$\left(P_c - P_b\right) \ / \ \left(1 - P_b\right)$ P_c is the complete agreement probability, and P_b represents the likelihood 'by chance'. Its range is (-1, 1).			

The measures of accuracy account for the model's overall accuracy.

Sensitivity (Recall) is the proportion of true-positive lung cancer cases the model correctly identifies. Specificity is the percentage of non-cancer cases correctly identified. Precision reflects the model's accuracy in making positive predictions. The F1 Score balances precision and recall. Kappa gauges agreement between the model's predictions and actual labels, accounting for chance. AUC evaluates the model's ability to discriminate between positive and negative classes; higher scores indicate stronger performance.

Model Performance Assessment in Various Folds

Fold 1 (Introduction of Model)

RF, ET, and AJAD were far ahead in the first fold with a mark of 97. Although the GB model has a high recall (1.00), it exhibits low specificity (0.25), indicating difficulty correctly classifying non-cancer cases. The AdaBoost model demonstrated slightly reduced accuracy of 94, with balanced performance across all measures, such as sensitivity and specificity (1.00 and 0.50, respectively). As anticipated, the LR had low specificity (0.00), which reduced overall performance, although it was high on recall (1.00).

Fold 2 to Fold 4 (Constant Performance Patterns)

RF, ET, and AJAD maintained strong results in folds two to four, posting consistently high scores of 90-97%. These models sustained equal sensitivity and specificity. ET was especially stable, without loss of sensitivity or precision. GB's specificity remained weak (0.25-0.50), resulting in fluctuating performance and more false positives.

AdaBoost performed reliably across metrics, achieving 87-94%. LR improved in fold 2 with a sensitivity of 0.96 but lagged in specificity and Kappa, a disadvantage on imbalanced datasets.

Fold 5 to Fold 6 (Gradual Decline in Performance of Some Models)

The RF, ET, and AJAD models continued to reign in these folds, achieving accuracy rates of 94-97% and remaining robust. GB and AdaBoost were also slightly improved, especially in fold 6, with GB recording a higher specificity (1.00). Nevertheless, the overall quality of GB remained lower than that of RF and ET.

LR, however, showed a significant reduction in sensitivity to 0.84, with a minor decrease in accuracy. This finding also highlights that the LR may not be the most appropriate model for the given dataset, particularly when predicting rare events, as is the case in lung cancer prediction.

Fold 7 to Fold 9 (Model's Approach Optimal Performance)

RF, ET, and AJAD models reached peak performance and showed high AUC scores (0.94-1.00) with low fold-to-fold variance by fold 7. AdaBoost maintained good precision, but slightly lower precision (0.93 in fold 7). GB was found to be superior, with consistent specificity of 0.50-1.00, and, again, LR was lower in all cases, particularly in negative cases.

RF and AJAD achieved almost perfect results in fold 9, with an accuracy of 97 percent, indicating they are not vulnerable to overfitting. GB was highly sensitive, whereas AdaBoost was more precise and better at recall (0.96 and 1.00, respectively).

Fold 10 (Final Results)

RF, ET, and AJAD achieved top results in the final fold, each reaching 97% accuracy. Their stability across folds suggests high reliability in lung cancer detection. ET's perfect sensitivity (1.00) demonstrates its particular strength in identifying cancer cases. GB's sensitivity (0.93) and precision dipped, lowering its overall performance.

AdaBoost improved specificity (1.00) and achieved a high AUC (0.99), confirming its ability to

distinguish cancer from non-cancer cases. LR struggled, with sensitivity at 0.33 and low precision.

Discussion: Performance Analysis Model

Overall, RF, ET, and AJAD performed excellently across folds, maintaining high accuracy, sensitivity, and precision. These models effectively distinguished cancerous from non-cancerous cases. ET excelled in sensitivity, signaling strong cancer case prediction with minimal overfitting.

While GB achieved high recall in some folds, its low specificity resulted in excess false positives. GB effectively separated classes compared to other models in the AJAD configuration, which mitigated some weaknesses.

AdaBoost achieved moderate accuracy, with consistently high precision and recall, but its sensitivity lagged behind other models, suggesting it missed some true positives.

Despite its popularity, LR performed poorly on these datasets due to low specificity, resulting in frequent misclassification of non-cancerous cases. Its performance fluctuated and was highly sensitive to class imbalance, notably in folds 6 and 7.

Table 6. Model Accuracy Across Folds in 10 folds(Oesophageal Dataset)

Fold	**Model**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F1 Score**	**Kappa**	**AUC**
1	Random Forest	1	1	1	1	1	1	1	1
1	Gradient Boosting	1	1	1	1	1	1	1	1
1	Extra Trees	1	1	1	1	1	1	1	1
1	Ada Boost	0.91	0.38	0.99	0.91	0.38	0.54	0.50	0.96
1	Logistic Regression	0.97	0.89	0.99	0.92	0.89	0.91	0.89	1
1	AJAD	1	1	1	1	1	1	1	1
2	Random Forest	1	1	1	1	1	1	1	1
2	Gradient Boosting	1	1	1	1	1	1	1	1
2	Extra Trees	1	1	1	1	1	1	1	1
2	Ada Boost	0.89	0.27	0.99	0.75	0.27	0.40	0.35	0.96
2	Logistic Regression	0.95	0.80	0.97	0.81	0.80	0.81	0.78	0.98
2	AJAD	1	1	1	1	1	1	1	1
3	Random Forest	1	1	1	1	1	1	1	1
3	Gradient Boosting	1	1	1	1	1	1	1	1
3	Extra Trees	1	1	1	1	1	1	1	1
3	Ada Boost	0.90	0.31	1	1	0.31	0.47	0.44	0.97
3	Logistic Regression	0.96	0.78	0.99	0.96	0.78	0.86	0.84	1
3	AJAD	1	1	1	1	1	1	1	1
4	Random Forest	1	1	1	1	1	1	1	1
4	Gradient Boosting	1	1	1	1	1	1	1	1
4	Extra Trees	1	1	1	1	1	1	1	1
4	Ada Boost	0.90	0.25	1	1	0.25	0.41	0.37	0.96

4	Logistic Regression	0.96	0.82	0.99	0.90	0.82	0.86	0.84	0.99
4	Ensemble	1	1	1	1	1	1	1	1
5	Random Forest	1	1	1	1	1	1	1	1
5	Gradient Boosting	1	1	1	1	1	1	1	1
5	Extra Trees	1	1	1	1	1	1	1	1
5	Ada Boost	0.91	0.35	1	0.95	0.35	0.51	0.47	0.97
5	Logistic Regression	0.97	0.87	0.99	0.91	0.87	0.89	0.87	0.99
5	AJAD	1	1	1	1	1	1	1	1
6	Random Forest	1	1	1	1	1	1	1	1
6	Gradient Boosting	1	1	1	1	1	1	1	1
6	Extra Trees	1	1	1	1	1	1	1	1
6	Ada Boost	0.90	0.33	0.99	0.86	0.33	0.48	0.44	0.96
6	Logistic Regression	0.90	0.48	0.97	0.68	0.48	0.57	0.51	0.94
6	AJAD	1	1	1	1	1	1	1	1
7	Random Forest	1	1	1	1	1	1	1	1
7	Gradient Boosting	1	1	1	1	1	1	1	1
7	Extra Trees	1	1	1	1	1	1	1	1
7	Ada Boost	0.92	0.43	1	0.96	0.43	0.559	0.555	0.97
7	Logistic Regression	0.99	1	0.99	0.92	1	0.96	0.95	1
7	Ensemble	1	1	1	1	1	1	1	1
8	Random Forest	1	1	1	1	1	1	1	1
8	Gradient Boosting	1	1	1	1	1	1	1	1
8	Extra Trees	1	1	1	1	1	1	1	1
8	Ada Boost	0.90	0.33	0.99	0.82	0.33	0.47	0.43	0.96
8	Logistic Regression	0.96	0.87	0.98	0.85	0.87	0.86	0.84	0.98
8	Ensemble	1	1	1	1	1	1	1	1
9	Random Forest	1	1	1	1	1	1	1	1
9	Gradient Boosting	1	1	1	1	1	1	1	1
9	Extra Trees	1	1	1	1	1	1	1	1
9	Ada Boost	0.90	0.33	0.99	0.86	0.33	0.48	0.44	0.95
9	Logistic Regr	0.95	0.78	0.97	0.82	0.	0.80	0.77	0.

	essio n					7 8			9 8
9	Ense mble	1	1	1	1	1	1	1	1
1 0	Ran dom Fore st	1	1	1	1	1	1	1	1
1 0	Grad ient Boos ting	1	1	1	1	1	1	1	1
1 0	Extr a Tree s	1	1	1	1	1	1	1	1
1 0	Ada Boos t	0. 91	0. 38	1	0. 95	0 . 3 8	0. 5 5	0. 5 1	0 . 9 7
1 0	Logi stic Regr essio n	0. 95	0. 78	0. 98	0. 88	0 . 7 8	0. 8 3	0. 8 0	0 . 9 9
1 0	Ense mble	1	1	1	1	1	1	1	1

Model Performance Evaluation 10 Folds.

A 10-fold cross-validation was used to compare the performance of six models: RF, GB, ET, AdaBoost, LR, and AJAD, a combination of multiple algorithms. The analysis was conducted using several measures, including Accuracy, Sensitivity (Recall), Specificity, Precision, F1 Score, Kappa, and AUC. These metrics determine classification correctness and the distinction between the positive and negative classes, summarized in Table 6.

Evaluation results show that the tree-based models (RF, GB, ET) and AJAD are consistently relevant across all 10 folds, achieving near-perfect results. Their scores for accuracy, sensitivity, specificity, and AUC were all high, indicating suitability. AJAD achieved the highest performance, with perfect scores (1) in nearly all metrics and folds.

AdaBoost and LR showed greater variability in performance, especially in accuracy and recall. While generally accurate, they struggled with positive cases (low recall), notably in earlier folds. AdaBoost's sensitivity and recall ranged from 0.25 to 0.43. LR also showed some variability, with sensitivity dropping in certain folds, but it remained accurate (0.90-0.97) in most cases.

General Observations: Cross-fold results highlight the power of tree-based models and AJAD, exemplified by perfect scores in Fold 1 for RF, GB, ET, and AJAD. AdaBoost, despite low sensitivity and recall (0.38), achieved an accuracy of 0.91, indicating better detection of negatives than positives. LR was more sensitive, with accuracy (0.97) and AUC (1) suggesting effective identification of positives.

RF, GB, and ET delivered their best performance in Fold 2. AdaBoost slightly increased accuracy (0.89), although its sensitivity and recall were low (both 0.27), suggesting it continued to struggle to identify true positives. The performance of LR was good, with an accuracy of 0.95 and an AUC of 0.98, making it a strong competitor, albeit not the best.

This pattern continued in subsequent folds; AdaBoost stayed weak in sensitivity and recall, while LR's performance was mostly good. RF, GB, ET, and AJAD maintained ideal results across all metrics, with AJAD demonstrating exceptional stability and consistency across folds.

Discussion of major metrics: Model evaluation used accuracy, sensitivity, precision, and F1 score. Tree-based and AJAD models were nearly perfect (1) for accuracy and sensitivity. In high-sensitivity activities, these models excelled, with LR showing moderate values (0.78-0.89) and AdaBoost performing poorly (0.25-0.43).

Specificity, which measures the capacity to distinguish the true negatives correctly, was perfect in most models except in AdaBoost, where there was a slight drop in subsequent folds. Precision was also high, with models such as RF, GB, and ET demonstrating good precision: when they said a case was positive, they were likely correct. The F1 Score, which balances precision and recall, was ideal for the tree-based models and AJAD but showed greater variability with AdaBoost and LR.

Finally, the AUC, which measures class distinction, was perfect (1) for both AJAD and tree-based models. AdaBoost and LR had AUC values that were slightly lower than those of the top performers.

Leading Performing Models and Insights: Based on the assessment, the AJAD model outperformed all other models, scoring perfectly across all measures. RF, GB, and ET (tree-based models) performed excellently and indicated that these models are highly effective at addressing the task at hand. Although AdaBoost and LR achieved fair accuracy, they showed significant issues with sensitivity and recall, especially on the first folds. This means these models may fail to handle certain properties of the dataset, such as class imbalance or noisy features.

For applications that demand robustness and generalization, the AJAD model is unquestionably the best, followed by tree-based models. Nevertheless, AdaBoost and LR can still be useful when computational resources are limited or the models are not as complex as they need to be, but additional tuning would be necessary to improve their sensitivity and overall performance.

Accuracy / Sensitivity / Specificity / Precision (mean [95% CI])

Model	Accuracy	Sensitivity	Specificity	Precision
AJAD	0.924 [0.887, 0.961]	0.960 [0.927, 0.993]	0.675 [0.505, 0.845]	0.951 [0.927, 0.975]
AdaBoost	0.887 [0.827, 0.947]	0.918 [0.858, 0.978]	0.675 [0.486, 0.864]	0.951 [0.923, 0.979]
ExtraTrees	0.914 [0.871, 0.957]	0.952 [0.915, 0.989]	0.650 [0.477, 0.823]	0.948 [0.924, 0.972]
GradientBoosting	0.907 [0.869, 0.945]	0.960 [0.927, 0.993]	0.550 [0.365, 0.735]	0.935 [0.910, 0.960]
LogisticRegression	0.879 [0.844, 0.914]	0.974 [0.946, 10]	0.233 [0.129, 0.337]	0.898 [0.882, 0.914]
RandomForest	0.916 [0.882, 0.950]	0.960 [0.932, 0.988]	0.617 [0.527, 0.707]	0.944 [0.932, 0.956]

F1 / Kappa / AUC (mean [95% CI])

Model	F1	Kappa	AUC
AJAD	0.955 [0.935, 0.975]	0.644 [0.490, 0.798]	0.926 [0.878, 0.974]
AdaBoost	0.932 [0.895, 0.969]	0.557 [0.380, 0.734]	0.923 [0.868, 0.978]
ExtraTrees	0.950 [0.925, 0.975]	0.608 [0.439, 0.777]	0.902 [0.834, 0.970]
GradientBoosting	0.947 [0.926, 0.968]	0.540 [0.369, 0.711]	0.907 [0.850, 0.964]
LogisticRegression	0.934 [0.914, 0.954]	0.284 [0.127, 0.441]	0.842 [0.756, 0.928]
RandomForest	0.952 [0.935, 0.969]	0.613 [0.482, 0.744]	0.926 [0.879, 0.973]

AJAD performed best overall in 10-fold cross-validation, demonstrating consistently high central estimates and narrow uncertainty bounds for key metrics. The average Accuracy of 0.924 (95% CI: 0.887-0.961) and F1-score of 0.955 (95% CI: 0.935-0.975) clearly affirm AJAD's high precision-to-recall ratio for positive classifications. Notably, AJAD achieved the top chance-corrected agreement, with a Kappa of 0.644 (95% CI: 0.490-0.798). This outcome is significant and warrants rigorous evaluation, as Kappa minimizes the risk of overfitting in contexts prone to class imbalance or predictive bias. Ultimately, the results provide strong evidence that AJAD's superiority is robust and reflects genuine model performance, rather than artifacts from specific fold distributions.

One of the main strengths of AJAD is the balance between its errors. In most cases, this balance is the difference between an accurate and a deployable model. AJAD had high Sensitivity (0.960), indicating the model can accurately detect positive cases. It also demonstrated the best mean Specificity (0.675), meaning it does not overfit to positives as much as linear baselines can. This balance is further confirmed by AJAD's highest discriminative ability (AUC = 0.926, 95% CI: 0.878-0.974). Its discriminative ability is comparable to the best-performing tree ensemble, while also offering better agreement and stability across folds. Together, these results show that AJAD is a strong and confident alternative to traditional ensemble learners. It achieves both high discrimination and more reliable class behavior.

Conclusion

Oesophageal and lung cancer are among the world's most challenging and deadly diseases due to late diagnosis, high mortality, and complex treatment. This research aims to improve early diagnosis and treatment using AI-based solutions. By applying ensemble learning, the study compares machine learning algorithms—RF, GB, ET, AdaBoost, and LR—to identify lung and oesophageal cancer. The goal is to evaluate the benefit of combining multiple classifiers into the AJAD model and using soft voting to boost performance.

AJAD consistently outperformed individual classifiers—RF, GB, ET, and AdaBoost—across all folds and achieved near-perfect accuracy, sensitivity, specificity, precision, recall, F1-score, and AUC. Soft voting combined the strengths of each classifier, reduced bias and variance, and created a more reliable prediction model. These results show that ensemble learning, especially the AJAD model, can improve lung and oesophageal

cancer detection and classification, leading to timely diagnosis and better patient outcomes.

Despite these developments, the study also noted limitations, particularly in AdaBoost and LR, which showed varying performance across sensitivity and recall. These models were successful in some folds but struggled to detect true positives, suggesting they need further tuning and optimization to reach their full potential. Conversely, the tree-based models, especially RF and ET, performed well across all assessment measures, making them highly applicable to cancer detection.

AI, ensemble learning, and multi-classifier approaches can significantly improve lung and oesophageal cancer diagnostics. Using AJAD enables healthcare providers and researchers to achieve higher predictive accuracy, stronger cancer-detection models, earlier diagnoses, tailored treatments, and increased patient survival.

Future Directions

This research advances AI use in cancer detection, but further study can improve the AJAD classifier and outcomes of cancer diagnosis and treatment.

- Increasing Dataset and Feature Variety:
- Future work should expand the dataset used in this study. Include more clinical and molecular features, such as genetic markers, protein expression, and imaging data, to build a detailed prediction model. Adding multimodal data will give a more complete view of each patient's status and may improve the model's performance.
- Combination of Imaging and Genomic Data:
- Fusion of machine learning, genomic, and imaging data may yield more accurate predictions of cancer detection. The development of AI-based imaging, especially in radiology and pathology, is very promising for improving diagnostic rates. Future studies are needed to determine how image classification systems based on deep learning methods (e.g., convolutional neural networks) can be coupled with AJAD to enhance the early detection of cancer and tailor the cancer treatment course depending on tumor properties.
- Better Model Calibration:
- AJAD performed well, but model calibration can be improved. Before aggregation, ensure each classifier is well-calibrated to improve final predictions. Use techniques like Platt scaling or isotonic regression to fine-tune model output and align predicted probabilities with true class probabilities.
- Handling Class Imbalance:
- Medical data often has a class imbalance, with more healthy cases than cancer cases. Future research should address class imbalance through advanced oversampling, undersampling, and cost-sensitive learning. This can increase sensitivity for rare positives without reducing specificity.
- The real-world Implementation and validation:
- To continue, it is paramount to confirm the effectiveness of AJAD and other AI-based models in a clinical environment. This means the model should be tested on external data from other hospitals or medical centers to assess its applicability and robustness. The prospective validation study, incorporating a wide range of patients, will help ensure the model can be effectively applied in practice.
- Algorithms of Personalized treatment are developed:
- Personalized treatment is key for cancer care beyond early detection. Apply decision support systems using AJAD and clinical or genetic data to help oncologists choose the best treatment plans. Future research should make treatment recommendation systems dynamic by integrating patient-specific data and treatment outcomes.
- Artificial Intelligence-based Prognostic Lung and Oesophageal Cancer:
- Future work should develop AI-based prognostic tools for lung and oesophageal cancer. These tools can predict cancer recurrence and long-term survival using clinical, demographic, and treatment data. Such tools help patients and providers manage long-term cancer care.

Lung and oesophageal cancer diagnosis and treatment will advance with AI and machine learning. Continued use of models such as AJAD and multimodal data will help researchers and clinicians develop more effective, timely diagnostic methods, thereby improving patient care and survival.

Data availability: We used two datasets for this article. Both datasets underlying this study are publicly available on the Kaggle website. These data were derived from publicly available sources. Links to these datasets are given below.

https://www.kaggle.com/datasets/akashnath29/lung-cancer-dataset

https://www.kaggle.com/datasets/abhinaba1biswas/esophageal-cancer-dataset

Competing Interests: The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Funding Information: Not Applicable

References

Alohali, M. A., Alqahtani, H., Ebad, S. A., Alotaibi, F. A., & Cho, J. (2025). Optimized deep learning approach for lung cancer detection using flying fox optimization and bidirectional generative adversarial networks. *PeerJ Computer Science*, *11*, e2853.

Bidzińska, J., & Szurowska, E. (2023). See lung cancer with an AI. *Cancers*, *15*(4), 1321.

Chan, S. Machine Learning-Based Prediction on Diagnosis and Severity of Lung Cancer. *SCHOLARLY REVIEW JOURNAL Учредители: Leadership & Innovation Lab*, (11).

Chaudhuri, A. K., Sinha, D., & Thyagaraj, K. S. (2018). Identification of the recurrence of breast cancer by discriminant analysis. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2* (pp. 519-532). Singapore: Springer Singapore.

Chaudhuri, A. K., Sinha, D., & Thyagaraj, K. S. (2018). Identification of the recurrence of breast cancer by discriminant analysis. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2* (pp. 519-532). Singapore: Springer Singapore.

Chaudhuri, A. K., Das, S., & Ray, A. (2023). A hybrid feature selection and stacked generalization model to detect breast cancer. In *Data-centric AI solutions and emerging technologies in the healthcare ecosystem* (pp. 165-183). CRC Press.

Das, S., Chaudhuri, A. K., Das, S., & Ghosh, P. (2025). Multistage feature selection and stacked generalization model for cancer detection. *Scientific Reports*, *15*(1), 38124.

Das, S., Chaudhuri, A. K., Paul, D., & Ghosh, P. (2026). Sequential Attribute Designator (SAD): A Novel Feature-Selection Framework for Pulmonary Disease Research. In *Next-Generation Bioinformatics for Pulmonary Disease Research* (pp. 413-436). IGI Global Scientific Publishing.

Diaz-Gay, M., Zhang, T., Hoang, P. H., Leduc, C., Baine, M. K., Travis, W. D., ... & Landi, M. T. (2025). The mutagenic forces shaping the genomes of lung cancer in never smokers. *Nature*, *644*(8075), 133-144.

Farshchiha, S., Asoudeh, S., Kuhshuri, M. S., Eisaei, M., Azadi, M., & Hesaraki, S. (2025). A Comprehensive Analysis on Machine Learning based Methods for Lung Cancer Level Classification. *Intelligence-Based Medicine*, 100309.

Ghosh, K., & Bhattacharjee, V. (2024). Lung cancer prediction: A performance analysis of machine learning classifiers. *International Journal of Statistics and Applied Mathematics*, *9*(5), 28-33.

Hua, P., Olofson, A., Farhadi, F., Hondelink, L., Tsongalis, G., Dragnev, K., ... & Hassanpour, S. (2025). Predicting targeted therapy resistance in non-small cell lung cancer using multimodal machine learning. *Journal of Thoracic Disease*, *17*(10), 8700-8714.

Jiang, W., Zhang, B., Xu, J., Xue, L., & Wang, L. (2025). Current status and perspectives of esophageal cancer: a comprehensive review. *Cancer Communications*, *45*(3), 281-331.

Langley, J. E., Sibley, D., Chiekwe, J., Keats, M. R., Snow, S., Purcell, J., ... & Wallace, A. (2025). Prehabilitation program for lung and esophageal cancers (boosting recovery and activity through early wellness): protocol for a nonrandomized trial. *JMIR Research Protocols*, *14*(1), e60791.

Singh, D. P. (2024). An extensive analysis of machine learning techniques for predicting the onset of lung cancer. *Tuijin Jishu/Journal of Propulsion Technology*, *45*(4), 2024.

Smolarz, B., Łukasiewicz, H., Samulak, D., Piekarska, E., Kołaciński, R., & Romanowicz, H. (2025). Lung cancer—epidemiology, pathogenesis, treatment and molecular aspect (review of literature). *International Journal of Molecular Sciences*, *26*(5), 2049.

van Tilburg, L., van de Ven, S. E., Spaander, M. C., van Kleef, L. A., Cornelissen, R., Bruno, M. J., & Koch, A. D. (2023). Prevalence of lung tumors in patients with esophageal squamous cell carcinoma and vice versa: a systematic review and meta-analysis. *Journal of Cancer Research and Clinical Oncology*, *149*(5), 1811-1823.

Yang, L., & Yang, F. (2025). Case Report: Esophageal malignant melanoma with lung adenocarcinoma: a rare case of dual primary cancers. *Frontiers in Oncology*, *15*, 1546806.

Zhong, J., Wang, Y., Zhu, D., & Wang, Z. (2025). A Narrative Review on Large AI Models in Lung Cancer Screening, Diagnosis, and Treatment Planning. *arXiv preprint arXiv:2506.07236*.