SUST Journal of Science and Technology Vol. 35(2) 2025, 12 - 27

SUST Journal of Science and Technology

journal homepage: https://journals.sust.edu

Pan-Cancer Classification in Clinical Applications: A Comprehensive Analysis using Machine Learning Models

Md. Shadmim Hasan Sifat *, Dr. Mohammad Shahidur Rahman

Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet – 3114, Bangladesh

ARTICLE INFO

Keywords:

Pan-cancer classification
Cancer type classification
Tissue of origin
Machine learning
Clinical data
XGBoost
Feature importance

ABSTRACT

Cancer affects millions of people worldwide, and accurate, timely classification of cancer type is closely tied to treatment selection and patient outcomes — particularly in metastatic disease or Cancer of Unknown Primary, where identifying the tissue of origin directly informs site-specific treatment. Multiple clinical and genomic variables, including mutation count, tumor stage, diagnosis age, tumor purity, smoking status and sex are individually associated with cancer behavior, but whether their combined use improves classification across a broad range of cancer types — rather than within a single cancer type, as in most prior work — has rarely been tested directly. This study evaluates four machine learning classifiers — Random Forest, Extreme Gradient Boosting, Support Vector Machine, and Artificial Neural Network — on a clinical pan-cancer dataset spanning 20 cancer types (6,110 patients; 1,222 held out for testing), using ten routinely available clinical and summary genomic covariates. XGBoost achieved the strongest performance (accuracy 87.0%, 95% CI: 85.0–88.8%; F1-score 0.869; ROC AUC 0.993), with Random Forest close behind; both outperformed SVM and ANN by a margin confirmed as statistically significant through 5-fold cross-validation. Rather than assuming that the selected variables jointly contribute to classification performance, this study tests that claim directly using permutation feature importance, finding that multiple features contribute distinct, non-redundant predictive value across all four models. This performance is broadly comparable to prior multi-class pan-cancer classification studies covering fewer cancer types (13–14) while relying on substantially higher-dimensional genomic inputs, suggesting that a smaller, more clinically accessible feature set can support classification across a wider range of cancer types. These results support the feasibility of clinical-data-driven pan-cancer classification as a direction for future decision-support research; external validation on independent, multi-institutional cohorts is required before any clinical application.

1. Introduction

Cancer remains one of the leading causes of death worldwide, and accurate, timely diagnosis is critical to improving patient outcomes [1]. Over the past decade, artificial intelligence (AI) and machine learning (ML) have become established tools in cancer research and clinical decision support, with algorithms capable of learning patterns from large volumes of clinical, imaging, and genomic data to assist in detection, diagnosis, and treatment planning [2, 3]. These approaches can process substantially more information than manual review alone — including histopathology imaging and structured patient records — to identify patterns associated with malignancy [4] and have grown more precise as both data availability and modeling techniques have matured, positioning ML as an increasingly practical component of clinical workflows rather than a purely experimental one [5].

Cancer, however, is not a single disease but a broad cat-

* Corresponding author.
E-mail address: shadmim-cse@sust.edu (Md. S. H. Sifat)

https://doi.org/10.63512/sustjst.2025.2002

Copyright© 2025 SUST. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

egory encompassing many biologically and clinically distinct malignancies, each with its own typical progression, treatment pathway, and diagnostic markers. This diversity presents a particular challenge in cases of metastatic disease or Cancer of Unknown Primary, where the tissue of origin is not immediately apparent from clinical presentation alone, yet is often essential for selecting an appropriate site-specific treatment — a clinical problem that has itself motivated prior ML-based tissue-of-origin classification work [6, 7]. Most existing ML-based cancer prediction models, as discussed in Section 2, are developed and validated for a single cancer type, addressing detection or prognosis within that type rather than distinguishing among many cancer types simultaneously.

1.1. Need of the Proposed System

Several clinical and genomic variables — including mutation count, patient age, smoking status, tumor purity, tumor stage, and sex — are individually established as relevant to cancer risk, progression, or outcome (Section 2.4). Whether these variables contribute distinct, complementary predictive information when used jointly, and whether a model built on this combination can distinguish among a broad range of cancer types rather than one, has not been directly tested in prior work at the scale addressed here. Existing multi-class pan-cancer classification approaches that do address this broader task typically rely on high-dimensional genomic inputs, such as whole-exome sequencing or large curated mutation panels, that require sequencing infrastructure not uniformly available outside specialized cancer centers (Section 2.3).

This study addresses that gap by evaluating four ML classifiers — Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Artificial Neural Network (ANN) on a combination of ten routinely available clinical and summary genomic covariates, applied to a pan-cancer clinical dataset spanning 20 cancer types. Rather than assuming these variables jointly improve classification performance, this study tests that claim directly through permutation feature importance analysis (Section 4.8), in addition to reporting classification performance across all four models with cross-validated confidence intervals and significance testing (Sections 4.3–4.5). The aim is to evaluate the feasibility of pan-cancer classification using a feature set considerably smaller and more clinically accessible than those used in prior multi-class approaches, with a broader view toward downstream applications such as tissue-of-origin identification in metastatic or Cancer of Unknown Primary cases. As with any model of this kind, its performance here reflects a single retrospective cohort; the results are presented as evidence of feasibility rather than a claim of clinical readiness, with the requirements for further validation discussed in Section 5.

2. Literature Review

2.1. Data Modalities in AI-Driven Cancer Research

Recent advances in health informatics have highlighted the growing role of ML and deep learning techniques in integrating multi-modal clinical data to support disease diagnosis, prognosis, and treatment planning [8]. The various ML models designed for diagnosing cancer include a range of different types of data, which give different aspects of the disease: imaging data (X-ray, CT, MRI, and PET) captures tumor localization and morphological and developmental patterns [9]; genomic data includes DNA sequencing, gene expression, and mutations relevant to the diagnostic process and therapy selection [10, 11]; electronic health records (EHR) data is used to capture demographic information and history of treatments for outcomes analysis and therapeutic effectiveness [12]; structured clinical data, including demographics, symptoms, treatment paths, lab results, and pathologies, helps to analyze disease development [13, 14]. The use of ensembles, like soft-voting classifiers combining multiple base classifiers, has shown better performance in the field of breast cancer diagnostics [15], being a trend of using model family integration rather than one particular classifier.

2.2. Clinical-Data-Based Prediction for Single Cancer Types

There have been multiple studies which have leveraged ML models for analyzing structured clinical/laboratory data specific to particular cancers. Gould et al. developed an XGBoost model, using a relatively large database of 6,505 Non-Small Cell Lung Cancer (NSCLC) cases and 189,597 controls, demonstrating that the model was able to detect lung cancer 9–12 months earlier than usual because it met the standard criteria for screening eligibility [16]. In their study, Botlagunta et al. used features related to hematological parameters (monocytes and lymphocytes) obtained from EHR to predict metastasis in patients with breast cancer using a decision tree classifier and found an accuracy of 83% [13]. Jiang and Xu compared performance of a deep feed forward neural network (DFNN) against 9 other ML methods (Naive Bayes, SVM, RF and 6 others), finding out that RF and XGBoost models worked better when there were more samples belonging to one class (imbalanced data), while DFNN, SVM, and logistic regression (LR) models worked better on balanced data [17]. Chen et al. used XGBoost to analyze 18,220 clinical features obtained from EHRs to recognize early-stage pancreatic cancer, providing a ROC AUC of 0.84, 60% sensitivity, 90% specificity, and detection capability of up to 24 months before the disease [18]. Ahamad et al. used RF, SVM, and XGBoost algorithms to analyze clinical data from 349 patients to separate benign and malignant ovarian tumors, identifying CA125, CA19-9, and HE4 as biomarkers, and achieved 91% classification accuracy [19]. Okagbue et al. applied several classifiers, including AdaBoost, to 300 patient records from a single Nigerian hospital to predict breast cancer survival, with AdaBoost reaching 98.3% accuracy and a 99.9% ROC AUC [20]. Genetic-algorithm-optimized SVM and K-Nearest Neighbors models have also been applied to breast cancer detection in resource-constrained settings, improving on baseline classifier accuracy without requiring additional data collection [21].

These studies collectively demonstrate that clinical and laboratory data support effective ML-based prediction, but three limitations recur across them and constrain what can be concluded from their reported metrics. First, nearly all

frame the task as binary or few-class classification (e.g., malignant vs. benign, metastatic vs. non-metastatic, alive vs. dead) rather than distinguishing among many cancer types simultaneously; classification difficulty scales sharply with the number of target classes; a chance-level baseline for a balanced binary task is 50%, versus 5% for a balanced 20-class task, so accuracy figures from these studies are not directly comparable to a multi-class pan-cancer accuracy figure without accounting for this difference. Second, sample sizes vary by three orders of magnitude across these studies (300–349 patients in [19, 20] versus 196,000+ in [16]), and the highest-reported metrics — such as the 98.3% accuracy and 99.9% ROC AUC in [20] — come from the smaller, single-institution cohorts, where such near-perfect performance is difficult to distinguish from overfitting to a narrow patient population without external validation. Third, none of these studies test multiple clinical or genomic covariates jointly across more than a small number of cancer types, leaving open whether the same modeling approach generalizes as class diversity increases.

2.3. Multi-Cancer Classification Using Genomic Data

A smaller set of studies addresses multi-class cancer-type classification directly, though using higher-dimensional genomic inputs than routine clinical data. He et al. developed a random forest framework using a 600-gene somatic-mutation panel to trace tumor tissue-of-origin across 4,909 samples spanning 13 cancer types, achieving 88.2% accuracy and an F1-score of 88.9% under 10-fold cross-validation [7]. The first is Li and Luo, who developed a performance-weighted ensemble comprising of LR, SVM, RF, XGBoost, and ANN for cancer-type classification with 6,249 samples including 14 different cancer types, concluding that their weighted ensemble outperformed each single classifier [6].

The two above-described works are most similar to our current work and show the possibility of multi-class cancer-type classification using machine learning approaches with a large number of classes (up to 14). Both works use high dimensional molecular features – whole exome sequencing mutation calls or curated gene sets, comprising of several hundreds of genes, which require considerable sequencing infrastructure and expense and are not universally available even at major cancer genomics centers. Neither work uses non-molecular (clinical) variables like patient age, smoking status, tumor stage, sex in combination with mutation-based features in a unified model. Neither work examines specific variables used as input in the multi-class classification procedure, just providing aggregated accuracy results. A side-by-side comparison of these studies with the single-cancer studies reviewed in Section 2.2 and with the present work, including data type, number of classes, sample size, and reported metrics, is provided in Table 1.

2.4. Individual Predictor Variables in Cancer Outcome Modeling

Several of the variables used in the present study have established, if individually studied, associations with cancer behavior. Mutation count has been linked to cancer aggressiveness and genetic risk profiling [6, 7, 22]. Smoking status is an established predictor of both cancer risk and mortality outcomes [23, 24], and patient age is consistently associated with cancer incidence and prognosis [20, 25]. Tumor purity — the fraction of cancerous cells in a sampled specimen — affects the reliability of molecular profiling and has itself been the subject of dedicated prediction models [26, 27]. Sex-based differences in cancer incidence and progression, driven by hormonal, genetic, and environmental factors, are documented across multiple cancer types [28, 29]. In each case, however, the variable is typically examined in isolation or within a single cancer type, rather than jointly with the others across a broad multi-cancer panel; where multiple such variables do appear together as inputs to a single model (e.g., in several of the single-cancer studies above), their individual and combined contribution to model performance is rarely tested directly.

2.5. Research Gap and Motivation

Three gaps follow from the literature reviewed above. First, while mutation count, patient age, smoking status, tumor purity, and sex are each individually established as relevant to cancer behavior, their combined and interacting contribution has not been tested jointly across a broad panel of cancer types, nor directly measured through feature-level analysis rather than assumed from inclusion in a model's input set. Second, the two studies that do address multi-class cancer-type classification directly [6, 7] rely on high-dimensional whole-exome or curated gene-panel mutation data spanning 13–14 cancer types; whether a smaller set of routinely available clinical and summary genomic covariates can support classification across a broader 20-type panel has not been established. Third, the majority of clinical-data-based cancer prediction studies address binary or few-class problems, leaving limited prior evidence for how such approaches perform as the number of target classes scales toward the higher end of clinically encountered cancer diversity.

This study addresses these gaps directly: it evaluates four ML classifiers (RF, XGBoost, SVM, ANN) on a combined set of clinical and summary genomic covariates — mutation count, diagnosis age, DNA input, sample coverage, tumor purity, tumor stage, sex, smoking status, specimen type, and sample type — across 20 cancer types drawn from a single pan-cancer clinical cohort, and tests the combined-effect motivation directly through permutation feature importance (Section 4.8) rather than asserting it from model inclusion alone. This is presented as an empirical evaluation of feasibility and relative model performance under these conditions, not as a claim of clinical readiness; external validation on independent cohorts, discussed in Section 5, would be required before any such conclusion.

3. Materials and Methods

The current analysis employs clinical data obtained from the China Pan-Cancer (OrigiMed, Nature 2022) dataset available via cBioPortal at https://www.cbioportal.org/study/clinicalData?id=pan_origimed_2020. The OrigiMed study examines genomic alterations of 10,194 solid tumors in a Chinese population cohort. The choice of this particular dataset has been made with a goal

Table 1. Comparison of Representative Prior Studies

Study	Data Type	Cancer Type(s)	Classes	Model(s)	N	Key Reported Metric	Comparability Note
Gould et al. [16]	Clinical/lab	NSCLC (screening)	2 (binary)	XGBoost	196,102	Outperformed screening criteria	Binary; huge N; not multi-class
Botlagunta et al. [13]	EHR/hematology	Breast (metastasis)	2 (binary)	Decision Tree	Not stated	83% accuracy	Binary task
Jiang & Xu [17]	Clinical	Breast (metastasis)	2 (binary)	DFNN + 9 ML models	15-yr cohort	RF/XGBoost best on imbalanced data	Binary; no single headline accuracy reported
Chen et al. [18]	EHR (18,220 feat.)	Pancreatic	2 (binary)	XGBoost	Not stated	AUC 0.84, sens 60%, spec 90%	Binary; high feature count, small effective signal (low sensitivity)
Ahamad et al. [19]	Clinical/biomarkers	Ovarian	2 (binary)	RF, SVM, XGBoost	349	91% accuracy	Binary; small N
Okagbue et al. [20]	Clinical	Breast (survival)	2 (binary)	AdaBoost (best of several)	300	98.3% accuracy, 99.9% AUC	Binary; small, single-institution N — near-perfect metrics warrant caution
Li & Luo [6]	Whole-exome mutation	Pan-cancer	14 (multi-class)	5-classifier weighted ensemble	6,249	Best of compared classifiers (exact % not reported)	Closest prior multi-class comparator; high-dimensional genomic input only
He et al. [7]	Somatic mutation (600-gene panel)	Pan-cancer	13 (multi-class)	RF	4,909	88.2% accuracy, 88.9% F1 (10-fold CV)	Closest prior multi-class comparator; high-dimensional genomic input only
This study	**Clinical + limited genomic covariates**	**Pan-cancer**	**20 (multi-class)**	**RF, XGBoost, SVM, ANN**	**6,110 (1,222 test)**	**87.0% accuracy, 0.869 F1 (XGBoost, 95% CI 0.850–0.888)**	**Broadest class count among multi-class comparators; uses routinely available covariates rather than large-scale sequencing panels**

to examine how clinical and genomic factors together influence the progression of different solid tumors. The cancer types along with the number of corresponding samples are presented in Table 2.

This dataset contains a number of clinical and genomic variables relevant to the problem of cancer classification. These variables are: Mutation Count, Diagnosis Age, DNA Input, Sample Coverage, Sex, Smoke Status, Tumor Purity, Tumor Stage, Specimen Type, and Sample Type. The Cancer Type variable is the target of classification but not an input variable. There are 25 cancer types in the source dataset; however, five types of cancers (Colorectal Carcinoma, Esophageal Carcinoma, Pancreatic Cancer, Gastric Cancer, and Intrahepatic Cholangiocarcinoma) were removed from the dataset for data quality reasons. The abnormalities in the key clinical variables observed during the preprocessing stage of these five cancer types make the training of the machine learning models impossible. As a result, 20 cancer types have been used in this analysis, which are presented in Table 2. Excluding the mentioned five cancer types (3,689 patients) and 395 patients who have no data on Mutation Count, the resulting analytic sample included 6,110 patients from 20 cancer types, i.e., decreased by 40.1% from the initial 10,194 patients.

Table 2. Cancer Types and their Sample Counts from Dataset

Cancer Type	Sample Count
Non-Small Cell Lung Cancer	2000
Liver Hepatocellular Carcinoma	1122
Soft Tissue Sarcoma	448
Extrahepatic Cholangiocarcinoma	339
Breast Carcinoma	313
Kidney Renal Cell Carcinoma	282
Ovarian Carcinoma	252
Gallbladder Carcinoma	239
Small Cell Lung Cancer	214
Head and Neck Carcinoma	150
Bone Sarcoma	144
Cancer of Unknown Primary	116
Carcinoma of Uterine Cervix	102
Urothelial Carcinoma	94
Gastrointestinal Neuroendocrine Tumor	71
Small Bowel Carcinoma	58
Uterine Corpus Endometrial Carcinoma	57
Melanoma	53
Thymic Tumor	28
Thyroid Carcinoma	28

3.1. Data Processing

3.1.1. Data Filtering

During preprocessing, missing values were identified in the Mutation Count field for 395 patients. Rows with missing values for this feature were removed rather than imputed, to avoid introducing potential bias or inaccuracy through estimated values for a variable central to the analysis.

3.1.2. Data Encoding

As previously mentioned, the dataset includes features that are classified as either discrete or continuous. Continuous features that were utilized straight for their analysis include Sample Coverage, Diagnosis Age, DNA Input, Mutation Count, and Tumor Purity. Categorical features that are discrete, such as Tumor Stage, Sample Type, Sex, Smoke Status, Specimen Type and Cancer Type were required to be encoded for analysis. The 'LabelEncoder' function from the 'sklearn.preprocessing' package was utilized to transform categorical features into numerical equivalents. Each unique category was preserved and mapped to an integer encoded as a 'Fit transform' function. The encoded category was preserved by the original category's identity through the value it was mapped to which was numerically defined, and which was encapsulated in a dictionary. The encoding process improved interpretability by allowing straightforward conversion of encoded data back to its original form when necessary.

3.1.3. Data Splitting Strategy

The dataset was partitioned using a stratified train/validation/test split to preserve the relative proportions of all 20 cancer types across subsets. An initial split allocated 20% of the data to a held-out test set, with the remaining 80% retained for model development. From

this 80%, a further 10% was set aside as a validation set for hyperparameter tuning and model selection, leaving approximately 72% of the total dataset for training, 8% for validation, and 20% for testing. Stratified sampling was used at both splitting steps to ensure that even the rarest classes (e.g., Thymic Tumor, Thyroid Carcinoma) were represented proportionally in each subset. The test and validation sets were held out prior to any resampling or feature scaling and were not used to fit any transformation applied to the training data.

3.1.4. Data Balancing

Class imbalance was addressed within the training fold only, using a combination of undersampling and oversampling applied strictly after the train/validation/test split. Classes with more than 1,000 training samples were downsampled to 1,000 using random undersampling. Classes with fewer than 1,000 training samples were oversampled using the Synthetic Minority Oversampling Technique (SMOTE) [30], with each class's oversampling target capped at 10 times its original sample count (and not exceeding 1,000), rather than a fixed target shared across all classes. This ratio-based cap bounds the proportion of synthetic samples consistently across classes of very different sizes: a class with several hundred real samples and a class with only a few dozen are both oversampled by the same multiplicative factor rather than being pulled toward the same fixed count, which would otherwise leave the smallest classes composed almost entirely of synthetic data. The number of neighbors used by SMOTE was set adaptively per class, bounded by the number of real samples available in the smallest oversampled class, to avoid degenerate behavior when a class has very few original examples. We note that even under a 10x cap, the smallest classes (approximately 28–40 original samples) remain predominantly synthetic after resampling, which is discussed as a limitation in Section 5.3.

Both undersampling and oversampling were implemented as steps within a single scikit-learn/imbalanced-learn pipeline alongside feature scaling and the classifier, ensuring that resampling was refit independently within each cross-validation fold and never applied to the validation or test sets.

3.1.5. Data Standardization & Normalization

Continuous features were standardized to zero mean and unit variance using scikit-learn's `StandardScaler`. The scaler was fit exclusively on the training set, and the resulting mean and variance were then used to transform the validation and test sets, ensuring all subsets were scaled consistently without the test data influencing the fitted parameters. Standardization works especially well for models that are sensitive to the ratio of different inputs — such as the SVM model and the ANN model that we use in this research, which are distance-based and gradient descent–optimized, respectively. On the other hand, decision tree ensembles like RF and XGBoost are invariant to monotonic transformations of the input features and have been added to the pipeline only for the sake of uniformity.

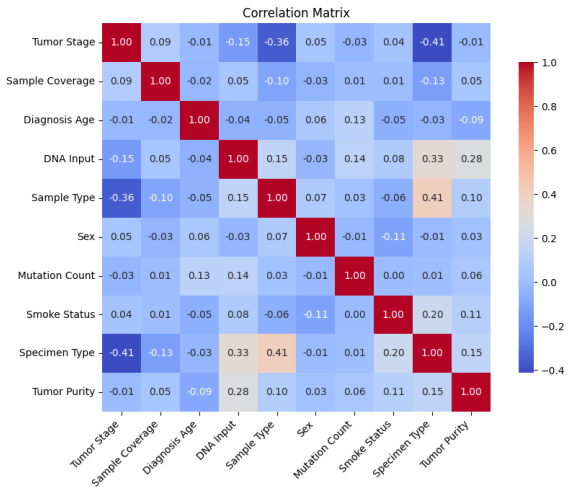


Figure 1. Correlation matrix for selected features.

3.2. Correlation Analysis

A correlation analysis was conducted on the ten clinical and genomic features to assess pairwise linear association (Figure 1).

The large majority of feature pairs showed near-zero correlation ($|r| < 0.15$), indicating that most of the selected features carry largely independent, non-redundant information. Specimen Type showed somewhat higher correlation with a few other variables — Tumor Stage ($r = -0.41$), Sample Type ($r = 0.41$), and DNA Input ($r = 0.33$) — with weaker associations to Smoke Status ($r = 0.20$) and Tumor Purity ($r = 0.15$); DNA Input and Tumor Purity were also mildly correlated with each other ($r = 0.28$). These values remain well below the range typically associated with multicollinearity concerns (commonly $|r| > 0.7$–0.8), so no feature pair was considered redundant enough to warrant removal. This pattern is broadly consistent with the permutation importance results in Section 4.8, where multiple features, including several with only weak pairwise correlation to one another, each contribute distinct predictive value across all four models; a correlation analysis alone cannot establish which features are most predictive or how they interact, since it captures only pairwise linear relationships rather than each feature's contribution to classification performance, which Section 4.8 addresses directly.

3.3. ML Classifiers

ML approaches to analyzing patient data are broadly grouped into supervised and unsupervised learning [31, 32, 33]. Healthcare institutions routinely collect large volumes of de-identified patient data, which provides the historical basis on which classification models are developed and trained. A range of ML methods has been applied to cancer identification using both structured numerical data and medical imaging. This study compares four such classifiers directly on the same clinical dataset, evaluated on both training and held-out test performance.

3.3.1. Random Forest

RF, extensively used in clinical studies [34], is a supervised ensemble algorithm of multiple decision trees, each

Table 3. Parameters Considered for RF, XGBoost, SVM, and ANN Models

Model	Parameters
RF	Number of trees = 500 Minimum size of terminal nodes = 1 Number of features to be analyzed = $\sqrt{p}$, where p is the no. of features
XGBoost	Number of classes = 20 Maximum depth of each decision tree = 5 Learning rate = 0.1 Number of decision trees built sequentially = 1200 Fraction of the training data used for each boosting round = 0.7
SVM	Regularization parameter, C = 100 Influence of a single training example (gamma) = 1 Kernel function = Radial Basis Function (RBF) Decision function shape = 'ovr' ("one-vs-rest")
ANN	Number of Hidden Layers = 5 Number of Dropout Layers = 3 Drop-Out Amount per Dropout Layer = 0.02 Batch size = 32, Epochs = 100 Optimizer = Adam Loss Function = Sparse Categorical Cross-entropy Hidden layers activation function = ReLU Output layer activation function = Softmax Early Stopping: monitor = val_loss, patience = 8, min_delta = 0.005

with hierarchical nodes for feature conditions leading to classifications. It handles high-dimensional data, missing values, and various attribute types [35], but has high memory complexity for large datasets due to many trees [36]. It's more stable and noise-resistant than single trees [37]. Parameters are listed in Table 3.

3.3.2. Extreme Gradient Boosting

XGBoost is an algorithm for gradient boosting that creates sequential trees to fix previous mistakes and reduce losses [38, 39]. If it is improved by regularization to prevent overfitting and parallel tree building and efficient missing value management [40, 41], it shows its popularity among competitions with structured data [42]. The parameters used are shown in Table 3.

3.3.3. Support Vector Machine

SVMs separate classes of objects by mapping the data into a high-dimensional space with the use of kernels to identify a hyperplane that will maximize the margin between them [43]. However, SVMs are computationally expensive for huge datasets and tend to work poorly in a noisy environment [31]. The parameters used are presented in Table 3.

3.3.4. Artificial Neural Network

ANN resembles a biological neuron network which converts inputs into outputs [44]. Trained on minimizing the difference between predicted and actual values by backpropagation [45], ANNs efficiently work with complex, nonlinear patterns and are used in tasks like image classification, natural language processing, time series forecasting [46]. Proper design, tuning, and regularization are essential to avoid overfitting [47]. For the ANN, we implemented a fully connected feedforward network with one input layer, five hidden layers, and one output layer. The five hidden layers contain 64, 128, 256, 128, and 64 neurons respectively, each using ReLU activation. Dropout regularization was applied after the second, third, and fourth hidden layers (three dropout layers in total), while the first and fifth hidden layers were not followed by dropout. The output layer uses a softmax activation with 20 units, corresponding to the 20 cancer type classes. The key features were the input to the classifier, while the number of outputs to the predicted neurons was the predicted class. The modified version of stochastic gradient descent that we used for optimization was the Adam algorithm. It was used for assigning the parameters that minimized the loss function (sparse categorical cross-entropy) to the extent possible. We have used an "early stopping" function in Keras that stops the training process if overfitting can begin to occur (min delta = 0.005, patience = 8). The neural network was trained with the training set, with a separate validation set used during training for assessing hyper parameters of the model and preventing the model from overfitting. After the training is done, the model input in the testing set is used to evaluate the accuracy, sensitivity, specificity, and other metrics. It was developed in Python using the Keras package (3.4.1).

3.4. Evaluation Metrics

Since the dataset contains multiple cancer types with an imbalanced class distribution, different averaging strategies can be used to summarize classification performance. Macro averaging computes each metric independently for every class and then calculates their unweighted mean, assigning equal importance to all cancer types regardless of class frequency. Micro averaging aggregates the contributions of all classes before computing the metric and is equivalent to overall accuracy for single-label multiclass classification. Weighted averaging computes the metric for each class and averages the results according to the number of samples in each class, thereby accounting for class imbalance.

In this study, the reported overall precision, sensitivity (recall), F1-score, specificity, and ROC AUC were computed using weighted averaging, selected because it reflects the unequal distribution of samples across the 20 cancer classes rather than treating rare and common classes as equally weighted. Because specificity and ROC AUC are computed in a one-vs-rest manner for each class before averaging, both metrics benefit from the large number of true negatives available whenever one class is compared against the other 19 combined; this tends to push both metrics toward high values regardless of how well the model separates the 20 classes from each other, and they are interpreted with that in mind alongside precision, sensitivity, and F1 in Section 4.

3.5. Experimental Setup

All experiments were conducted in Python 3.10.12 on Google Colaboratory, using a CPU-only runtime with 12.7 GB RAM. Implementation used Scikit-learn, XGBoost, Imbalanced-learn, Keras/TensorFlow, Pandas, Matplotlib, and Seaborn. Approximate training times on the training set described in Section 3.1 were: RF, 30.9 seconds; XGBoost, 55.3 seconds; SVM, 69.7 seconds; and ANN, 74.8

Table 4. Comparison of Sensitivity and Specificity for RF, XGBoost, SVM, and ANN Models

Cancer Type	Sensitivity				Specificity			
	RF	XGB	SVM	ANN	RF	XGB	SVM	ANN
Bone Sarcoma	1.000	1.000	1.000	0.966	0.995	0.997	0.990	0.986
Breast Carcinoma	0.873	0.905	0.794	0.651	0.995	0.995	0.996	0.993
Cancer of Unknown Primary	1.000	1.000	1.000	1.000	0.990	0.996	0.997	0.992
Carcinoma of Uterine Cervix	1.000	1.000	1.000	1.000	0.993	0.995	0.994	0.991
Extrahepatic Cholangiocarcinoma	0.985	0.926	0.868	0.779	0.990	0.992	0.987	0.977
Gallbladder Carcinoma	1.000	1.000	0.958	0.771	0.995	0.993	0.991	0.991
Gastrointestinal Neuroendocrine Tumor	1.000	1.000	1.000	1.000	0.996	0.996	0.993	0.996
Head and Neck Carcinoma	1.000	1.000	1.000	1.000	0.982	0.990	0.977	0.971
Kidney Renal Cell Carcinoma	0.857	0.875	0.821	0.607	0.990	0.989	0.979	0.980
Liver Hepatocellular Carcinoma	0.723	0.759	0.567	0.616	0.987	0.989	0.988	0.979
Melanoma	1.000	1.000	1.000	1.000	0.995	0.997	0.994	0.995
Non-Small Cell Lung Cancer	0.780	0.833	0.772	0.748	0.995	0.989	0.976	0.985
Ovarian Carcinoma	1.000	0.960	0.920	0.820	0.990	0.990	0.991	0.985
Small Bowel Carcinoma	1.000	1.000	1.000	1.000	0.996	0.995	0.995	0.993
Small Cell Lung Cancer	1.000	1.000	1.000	0.767	0.980	0.989	0.974	0.975
Soft Tissue Sarcoma	0.778	0.800	0.711	0.611	0.991	0.985	0.981	0.973
Thymic Tumor	1.000	1.000	1.000	1.000	0.993	0.995	0.993	0.996
Thyroid Carcinoma	1.000	1.000	1.000	1.000	0.997	0.998	0.997	0.988
Urothelial Carcinoma	1.000	1.000	1.000	1.000	0.991	0.994	0.988	0.985
Uterine Corpus Endometrial Carcinoma	1.000	1.000	1.000	1.000	0.997	0.998	0.996	0.995

seconds over the full 100 training epochs (early stopping was configured but did not trigger within this run). RF was the fastest to train, consistent with its parallelizable, non-sequential tree construction, while XGBoost's sequential boosting incurred modest additional overhead. SVM's training time reflects its known $O(n^2)$–$O(n^3)$ complexity in the number of training samples, as noted in Section 3.3.3, though on this dataset size the ANN's per-epoch forward and backward passes accumulated the greatest total wall-clock cost across its 100 training epochs.

For each classifier, hyperparameters (Table 3) were selected through an iterative manual process: candidate values were adjusted and re-evaluated based on performance on the held-out validation set described in Section 3.1.3, with the objective of maximizing validation F1-score while monitoring for signs of overfitting. The test set was reserved exclusively for final performance reporting (Section 4) and was not used at any point during model selection or hyperparameter tuning. For each model, a small number of candidate values were tried around an order of magnitude below and above the final setting for each parameter, and the configuration yielding the best validation performance was retained (Table 3).

4. Results and Analysis

This section presents the classification performance of four models—ANN, RF, XGBoost, and SVM—trained to distinguish among 20 cancer types using clinical features, evaluated on a held-out test set across accuracy, precision, sensitivity, specificity, F1-score, log loss, ROC-AUC, 5-fold cross-validation, and confusion matrices. XGBoost and RF consistently outperform SVM and ANN across these measures, with differences shown to be statistically significant at the fold level; further external validation would be needed before drawing conclusions about clinical applicability.

4.1. Sensitivity and Specificity Analysis Among Classes

Sensitivity varied considerably across cancer types and models (Table 4), while specificity remained uniformly high throughout (0.971–0.998), consistent with the one-vs-rest averaging caveat discussed in Section 3.4. XGBoost achieved the highest sensitivity across most classes, followed closely by RF, with SVM and ANN trailing on the harder classes. Ten of the twenty classes — Cancer of Unknown Primary, Carcinoma of Uterine Cervix, Gastrointestinal Neuroendocrine Tumor, Head and Neck Carcinoma, Melanoma, Small Bowel Carcinoma, Thymic Tumor, Thyroid Carcinoma, Urothelial Carcinoma, and Uterine Corpus Endometrial Carcinoma — reached sensitivity of 1.000 under all four models; these are also among the smallest classes in the test set, so this should be read as an indication of feasibility rather than a strong claim of generalization, given how few test cases each estimate is based on. The weakest sensitivity was consistently observed for Liver Hepatocellular Carcinoma (0.567–0.759), Kidney Renal Cell Carcinoma (0.607–0.875), and Soft Tissue Sarcoma (0.611–0.800) across all four models, matching the confusion patterns identified in Section 4.7. Breast Carcinoma showed a widespread across models (0.651–0.905), and ANN was consistently the weakest of the four models on nearly every class where sensitivity fell below 1.0, including Small Cell Lung Cancer (0.767, versus 1.000 for the other three models) and Ovarian Carcinoma (0.820, versus 0.920–1.000 for the others). This pattern — ANN trailing the other three models on precisely the classes where any model struggles — is consistent with ANN's lower overall accuracy reported in Table 6.

4.2. Precision and F1-Score Analysis Among Classes

Precision was substantially more variable than sensitivity across classes, ranging from 0.263 (Thyroid Carcinoma, ANN) to 0.987 (Non-Small Cell Lung Cancer, RF) shown in

Table 5. Comparison of Precision and F1-Score for RF, XGBoost, SVM, and ANN Models

Cancer Type	Precision				F1-Score			
	RF	XGB	SVM	ANN	RF	XGB	SVM	ANN
Bone Sarcoma	0.829	0.879	0.707	0.622	0.906	0.935	0.829	0.757
Breast Carcinoma	0.902	0.905	0.909	0.837	0.887	0.905	0.847	0.732
Cancer of Unknown Primary	0.657	0.821	0.885	0.719	0.793	0.902	0.939	0.836
Carcinoma of Uterine Cervix	0.714	0.769	0.741	0.645	0.833	0.870	0.851	0.784
Extrahepatic Cholangiocarcinoma	0.859	0.875	0.797	0.671	0.918	0.900	0.831	0.721
Gallbladder Carcinoma	0.889	0.857	0.821	0.771	0.941	0.923	0.885	0.771
Gastrointestinal Neuroendocrine Tumor	0.737	0.737	0.636	0.737	0.848	0.848	0.778	0.848
Head and Neck Carcinoma	0.588	0.714	0.517	0.469	0.741	0.833	0.682	0.638
Kidney Renal Cell Carcinoma	0.800	0.790	0.657	0.596	0.828	0.831	0.730	0.602
Liver Hepatocellular Carcinoma	0.926	0.939	0.914	0.868	0.812	0.840	0.700	0.721
Melanoma	0.647	0.733	0.611	0.647	0.786	0.846	0.759	0.786
Non-Small Cell Lung Cancer	0.987	0.974	0.939	0.961	0.872	0.898	0.848	0.841
Ovarian Carcinoma	0.806	0.800	0.807	0.707	0.893	0.873	0.860	0.759
Small Bowel Carcinoma	0.688	0.647	0.647	0.550	0.815	0.786	0.786	0.710
Small Cell Lung Cancer	0.652	0.768	0.581	0.524	0.789	0.869	0.735	0.623
Soft Tissue Sarcoma	0.875	0.809	0.753	0.647	0.824	0.804	0.731	0.629
Thymic Tumor	0.429	0.500	0.429	0.545	0.600	0.667	0.600	0.706
Thyroid Carcinoma	0.556	0.625	0.556	0.263	0.714	0.769	0.714	0.417
Urothelial Carcinoma	0.633	0.731	0.559	0.514	0.776	0.844	0.717	0.679
Uterine Corpus Endometrial Carcinoma	0.750	0.800	0.706	0.667	0.857	0.889	0.828	0.800

Table 5. Notably, several of the classes with perfect sensitivity in Section 4.1 do not show correspondingly high precision: Thymic Tumor reaches sensitivity of 1.000 under every model but precision of only 0.429–0.545, and Head and Neck Carcinoma likewise reaches sensitivity of 1.000 with precision of only 0.469–0.714. This combination indicates that while these classes are rarely missed when they do occur, the models also misclassify a substantial number of other classes' cases as belonging to them — a pattern consistent with rare classes acting as an occasional default prediction rather than being cleanly separated from the rest of the feature space. F1-score reflects this trade-off directly: Thymic Tumor's F1 (0.600–0.706) is well below what its perfect sensitivity alone would suggest.

The opposite pattern holds for Non-Small Cell Lung Cancer, which combines high precision (0.939–0.987) with more moderate sensitivity (0.748–0.833, Section 4.1): when the models predict this class, they are rarely wrong, but they miss a meaningful share of true cases, consistent with the confusion toward Small Cell Lung Cancer and Head and Neck Carcinoma noted in Section 4.7. Across nearly every class, XGBoost achieved the highest or near-highest F1-score, RF followed closely, and ANN recorded the lowest F1-score of the four models on the majority of classes — most notably Thyroid Carcinoma, where ANN's F1 (0.417) trails RF, XGBoost, and SVM (0.714–0.769) by a wide margin, driven primarily by ANN's low precision (0.263) rather than its sensitivity, which was perfect for this class across all four models.

4.3. Classification Accuracy Across Four Models

Per-class classification accuracy was evaluated for each of the 20 cancer types across all four models, using the natural test-set distribution to reflect real-world class prevalence (Figure 2). Macro-averaged class accuracy was highest for XGBoost (0.870), followed by RF (0.847), SVM (0.792), and ANN (0.744), consistent with the ranking observed for

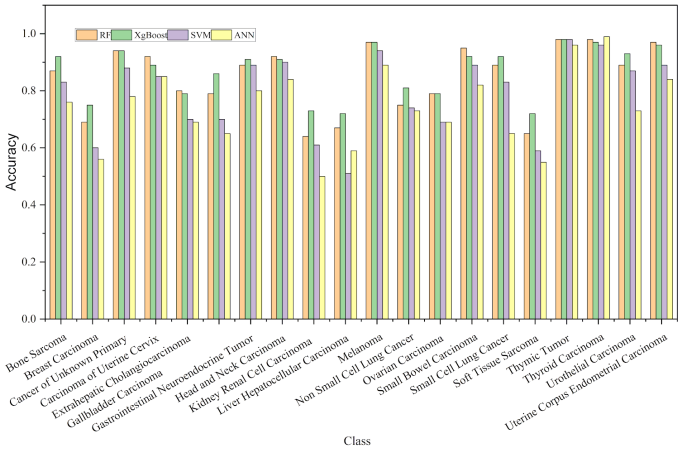


Figure 2. Class-wise accuracy comparison for RF, XGBoost, SVM, and ANN models.

overall accuracy, F1-score, and AUC (Section 4.5). Thymic Tumor, Thyroid Carcinoma, and Melanoma achieved the highest and most consistent accuracy across models (0.89–0.99), while Kidney Renal Cell Carcinoma, Liver Hepatocellular Carcinoma, Breast Carcinoma, and Soft Tissue Sarcoma were the most consistently difficult classes, with accuracy as low as 0.50–0.61 for the weaker models. This pattern closely mirrors the per-class AUC results (Section 4.6), with the same four cancer types ranking among the lowest-performing on both metrics regardless of model — evidence that these classes are genuinely harder to distinguish from the rest of the cohort, rather than the difficulty being an artifact of a single metric or model.

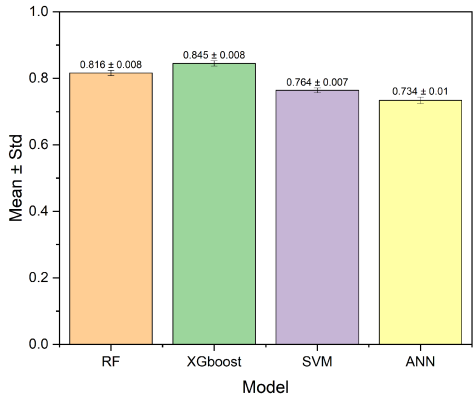


Figure 3. 5-fold cross-validation results for RF, XGBoost, SVM, and ANN models.

4.4. K-Fold Cross-Validation

Figure 3 reports 5-fold cross-validation accuracy (mean ± standard deviation) for each model: XGBoost 0.845 ± 0.008, RF 0.816 ± 0.008, SVM 0.764 ± 0.007, ANN 0.734 ± 0.010. Using a t-distribution with 4 degrees of freedom, the corresponding 95% confidence intervals are XGBoost (0.835, 0.855), RF (0.806, 0.826), SVM (0.755, 0.773), and ANN (0.722, 0.746). None of these intervals overlap, and pairwise Welch's t-tests confirm that every model differs significantly from each other at the fold level (all $p < 0.001$; e.g., XGBoost vs. RF: $t = 5.73$, $p < 0.001$; XGBoost vs. ANN: $t = 19.38$, $p < 0.001$). These p-values should be interpreted with some caution: because the five folds share overlapping training data, they are not strictly independent samples, which can cause standard t-tests to understate true uncertainty and yield smaller p-values than a fully independent-sample design would justify. The consistent, non-overlapping confidence intervals and stable per-fold performance nonetheless support the observed ranking — XGBoost > RF > SVM > ANN — as a robust pattern rather than an artifact of any single fold, even though the exact p-values above should be read as indicative rather than a fully rigorous test. The standard deviations are small and similar in magnitude across models (0.007–0.010), showing that each model's performance is stable from fold to fold rather than dependent on a favorable split. XGBoost's advantage over the other three models holds across every fold-level comparison, and its narrow interval, comparable to RF's, suggests the two ensemble methods are both more consistent across folds than SVM or ANN. The slightly wider interval for ANN (±0.010) is consistent with it being the most sensitive of the four models to which subset of data it is trained and evaluated on.

4.5. Overall Performance Comparison

Table 6 summarizes weighted-average performance across all 20 cancer types, with 95% confidence intervals for accuracy computed as Wilson score intervals given the test-set size. XGBoost achieved the best overall performance (accuracy 0.870, 95% CI: 0.850–0.888), with the lowest log loss (0.432) indicating the best-calibrated probability outputs among the four models. RF followed closely (accuracy 0.847, 95% CI: 0.826–0.866), with an interval that overlaps XGBoost's, indicating the two ensemble methods are not clearly separable on test-set accuracy alone (two-proportion z-test, $z = 1.63$, $p = 0.103$). SVM (0.794, 95% CI: 0.770–0.816) and ANN (0.745, 95% CI: 0.720–0.769) trail both ensemble methods by a margin that is statistically significant ($p < 0.01$ for every pairwise comparison involving SVM or ANN).

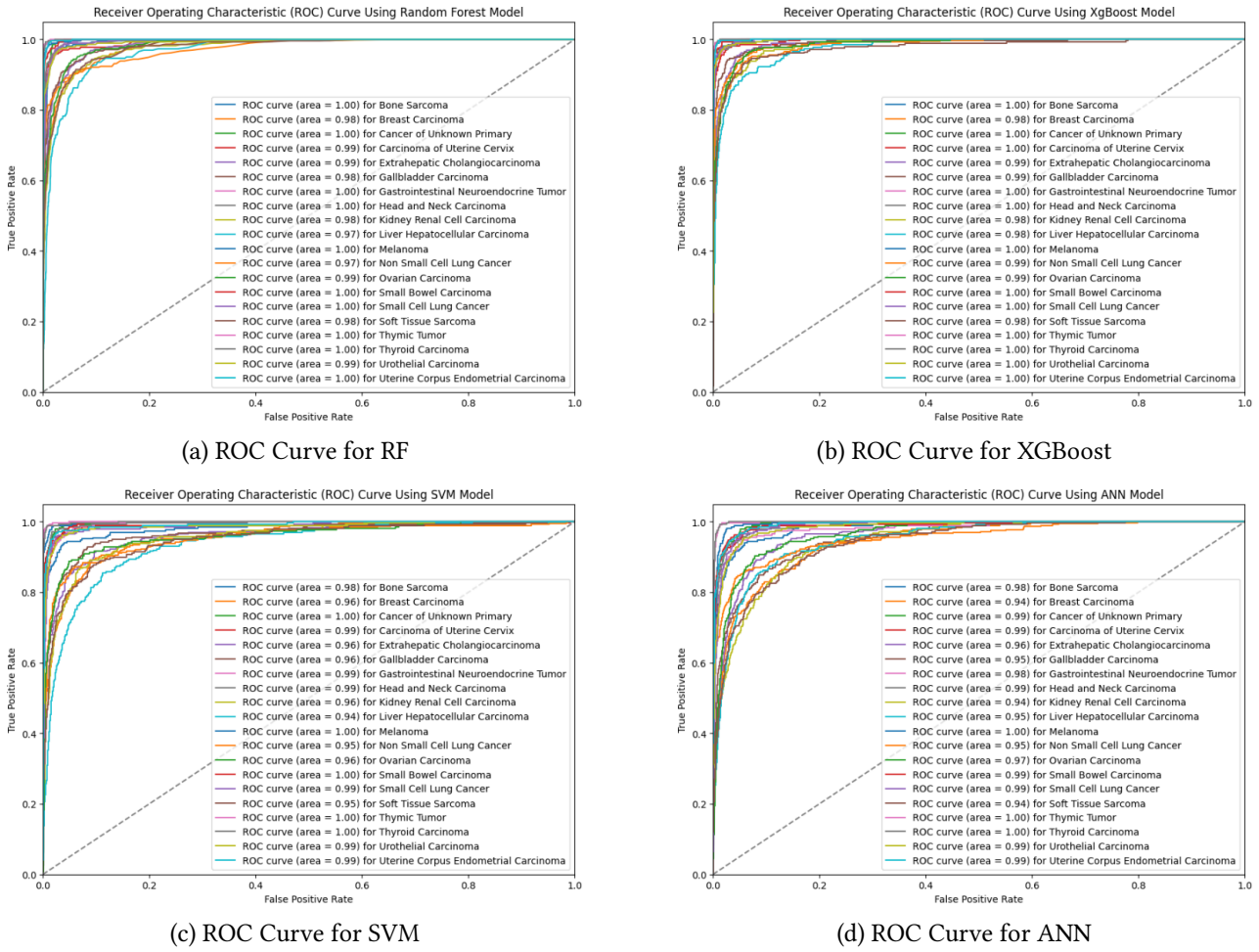
Table 6. Weighted-Average Evaluation of Models Across Classification Metrics

Metric	RF	XGB	SVM	ANN
Accuracy	0.847	0.870	0.794	0.745
95% CI	(0.826, 0.866)	(0.850, 0.888)	(0.770, 0.816)	(0.720, 0.769)
F1-Score	0.845	0.869	0.791	0.744
Precision	0.846	0.869	0.790	0.747
Specificity	0.992	0.993	0.989	0.986
Sensitivity	0.847	0.870	0.794	0.745
Log Loss	0.656	0.432	0.745	0.905
ROC AUC	0.990	0.993	0.977	0.975

Specificity is uniformly high across all four models (0.986–0.993). This reflects the one-vs-rest structure of the evaluation: for any given class, the negative set is the other 19 classes combined, which pushes specificity toward 1 regardless of how well the model separates the 20 classes from one another. ROC AUC (0.975–0.993) is subject to the same effect. We therefore treat accuracy, precision, sensitivity, and F1 as the primary basis for comparing models with specificity and ROC AUC reported for completeness. Weighting of all metrics is done using class support instead of macro-average weighting, which results in larger weight on the most common classes (Non-Small Cell Lung Cancer and Liver Hepatocellular Carcinoma) compared to less common classes. Test-set accuracy obtained for XGBoost (0.870) is slightly better than average 5-fold cross-validation (0.845, Section 4.4), but this difference is expected and not surprising. In case of 5-fold cross-validation, training is performed using four-fifths of the training subset (57.6% of the entire dataset) and in case of test-set accuracy, training is performed using the whole training subset (72% of the entire dataset), thus in the latter case 25% more training data is used. This is especially important in the case of very small and highly oversampled classes because an increase in the number of examples in each fold will have impact on performance. Cross-validation estimate given in Section 4.4 is a more conservative and reliable performance estimator since it uses five different training/test splits while the test-set accuracy uses only one split.

4.6. Class-Wise ROC AUC Curve Across Models

Discrimination capability was measured through ROC AUC curves in one-vs-rest format for each of the 20 cancer types (Figure 4). The best macro-averaged ROC AUC values were achieved with the use of XGBoost classifier (0.994), followed by RF (0.991), SVM (0.978), and ANN (0.975), which corresponds to the previous accuracy and F1 ranking results presented in Table 6. Per-class ROC AUC values



(a) ROC Curve for RF

(b) ROC Curve for XGBoost

(c) ROC Curve for SVM

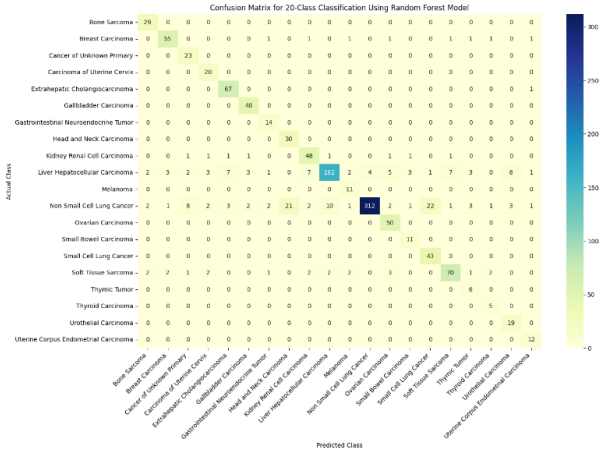
(d) ROC Curve for ANN

Figure 4. Receiver Operating Characteristic (ROC) curves for RF, XGBoost, SVM, and ANN models.

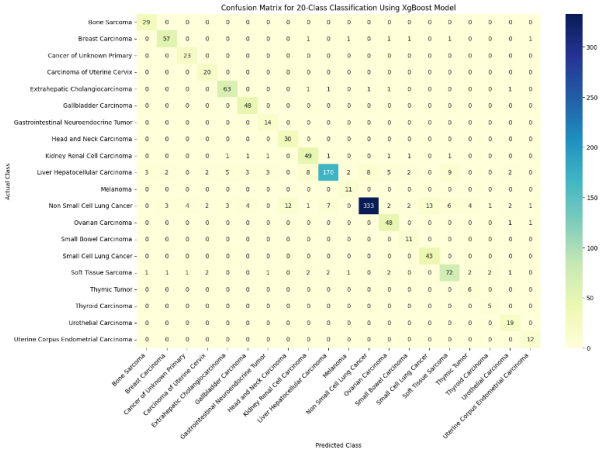
varied from 0.94 to 1.00 for all models with the most distinguished Melanoma, Thymic Tumor, and Thyroid Carcinoma being characterized with ROC AUC ≥ 0.99 with each applied classification method. On the other hand, the performance with Liver Hepatocellular Carcinoma, Kidney Renal Cell Carcinoma, Breast Carcinoma, and NSCLC was relatively lower (ROC AUC = 0.94–0.98), which indicates that such cancer types have more overlapping feature sets compared to other classes in the dataset and therefore can be more difficult to discriminate. It should be kept in mind when interpreting those ROC AUC values that they represent the one-vs-rest setting: with 20 classes, each per-class AUC measures discrimination of a given cancer type versus a pooled "everything else" class, which in turn consists of 19 different cancer types, quite distinct from each particular cancer type. This setup makes the classifier achieve relatively high ROC AUC values even for those classes that classifier separates relatively poorly because the negative class in this case can be well separated just by its unrelatedness to other cancer types. Therefore, relatively lower ROC AUC values of Liver Hepatocellular Carcinoma, Kidney Renal Cell Carcinoma, Breast Carcinoma, and NSCLC classes – which showed the biggest separation difficulties with all four classifiers – provide a more valuable indicator of discriminative problem compared to near-perfect ROC AUC values of well-separated Melanoma and Thymic Tumor classes.

4.7. Class-Wise Confusion Matrices Across Models

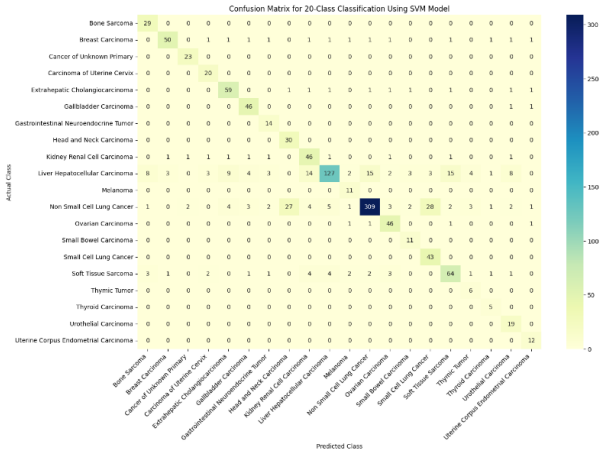
The confusion matrices in Figure 5 report classification outcomes for all 20 cancer types under RF, XGBoost, SVM, and ANN on the same held-out test set (n = 1,222). As regards accuracy of the classifiers, XGBoost showed the highest overall accuracy (87.0%), followed by RF (84.7%), SVM (79.4%) and ANN (74.5%). This is consistent with the order indicated in Table 6. In all four algorithms, the most common errors were always those of confusion between Non-Small Cell Lung Cancer and Small Cell Lung Cancer and also between Non-Small Cell Lung Cancer and Head and Neck Carcinoma. The former two types of cancer are anatomically/histologically close, thus a logical explanation for this overlap can be provided since the same set of features was used for classification. Another pattern was that of confusion between Liver Hepatocellular Carcinoma, Soft Tissue Sarcoma and Kidney Renal Cell Carcinoma in all four algorithms. Liver Hepatocellular Carcinoma showed the lowest recall of any high-support class, ranging from 0.567 (SVM) to 0.759 (XGBoost) depending on model, and Soft Tissue Sarcoma followed a similar pattern (0.611–0.800). Because these classes, together with Non-Small Cell Lung Cancer and Breast Carcinoma, make up a disproportionate share of the test set, they also account for most of the mod-



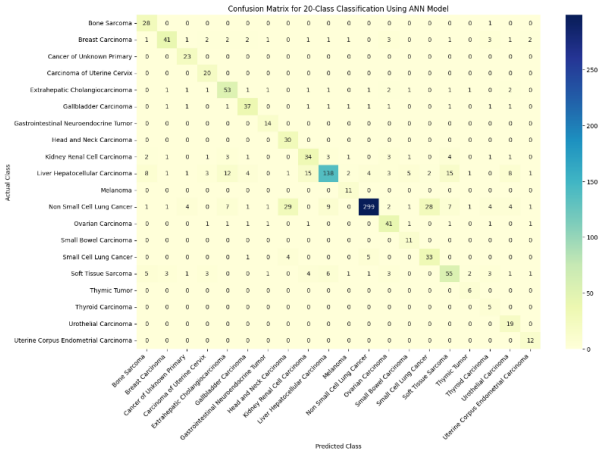
(a) Confusion Matrix for RF



(b) Confusion Matrix for XGBoost



(c) Confusion Matrix for SVM



(d) Confusion Matrix for ANN

Figure 5. Confusion matrices for RF, XGBoost, SVM, and ANN models.

els' absolute error counts; this is expected under class imbalance and should not be read as evidence that these cancers are intrinsically harder to separate than the rarer classes.

Non-Small Cell Lung Cancer, despite being the largest class by far ($n = 400$), was not among the weakest-performing classes: recall ranged from 0.748 (ANN) to 0.833 (XGBoost), roughly in line with each model's overall accuracy. Classes with very few test instances — Melanoma, Thymic Tumor, Thyroid Carcinoma, Urothelial Carcinoma, Small Bowel Carcinoma, and several others ($n = 5$–23 each) — reached recall of 1.0 under most or all four models. Given the small sample sizes involved, these results should be treated as indicative rather than conclusive, and we do not interpret them as evidence of superior model performance on these cancer types; a single additional misclassified case in any of these classes would produce a large swing in the reported recall.

Taken together, the matrices show that XGBoost reduces errors most consistently across the high-support, clinically overlapping classes (lung and hepato-renal-sarcoma groups) relative to RF, SVM, and ANN, with the gap between models widest on Liver Hepatocellular Carcinoma, Soft Tissue Sarcoma, and Kidney Renal Cell Carcinoma — the same classes identified as most difficult across every model.

4.8. Feature Importance Analysis

To test the combined-effect claim motivating this study, permutation importance was computed for all four models, measuring the drop in macro-F1 when each feature's values were randomly shuffled (10 repeats per feature; results in Table 7).

Four features — Mutation Count, DNA Input, Diagnosis Age, and Tumor Purity — ranked among the five most important predictors across all four models, indicating that these clinical and genomic variables consistently contribute non-redundant predictive information regardless of model architecture. Tumor Stage showed strong importance for SVM and ANN (ranking first for both, ΔF1 = 0.365 and 0.363 respectively) and moderate importance for RF (ΔF1 = 0.227) but contributed comparatively less to XGBoost (ΔF1 = 0.171). Sex and Smoke Status were the two least important features across all four models (ΔF1 ranging from 0.096–0.188), suggesting these variables carry the least independent predictive signal for cancer type classification within this feature set.

Notably, DNA Input — a sequencing-related technical covariate rather than one of the clinically motivated variables — was the single most important feature for XGBoost (ΔF1 = 0.395) and ranked among the top five for all other models. While this may reflect genuine biological signal (e.g.,

Table 7. Permutation Feature Importance (ΔMacro-F1) Across Models

Feature	RF	XGB	SVM	ANN
Mutation Count	0.288	0.282	0.212	0.211
DNA Input	0.261	0.395	0.331	0.332
Diagnosis Age	0.237	0.226	0.350	0.345
Tumor Stage	0.227	0.171	0.365	0.363
Tumor Purity	0.223	0.234	0.358	0.357
Smoke Status	0.188	0.171	0.164	0.165
Specimen Type	0.163	0.057	0.179	0.180
Sex	0.163	0.096	0.165	0.163
Sample Coverage	0.160	0.238	0.349	0.348
Sample Type	0.152	0.077	0.170	0.170

DNA input correlating with tumor cellularity, which varies by cancer type), it may also partly reflect systematic differences in sequencing protocol across cancer types in the source cohort; this possibility is discussed further as a limitation in Section 5.3.

A further pattern emerged between model families: RF and XGBoost consistently ranked Mutation Count and DNA Input as most important, while SVM and ANN showed a near-identical ranking prioritizing Tumor Stage, Tumor Purity, Sample Coverage, and Diagnosis Age above Mutation Count. This suggests tree-based ensembles and margin/gradient-based models weigh the same underlying feature set differently, rather than converging on an identical notion of feature relevance — itself a relevant methodological observation for future pan-cancer classification work.

Taken together, these results demonstrate — rather than assume — that multiple clinical and genomic features contribute distinct, non-overlapping predictive value across model types, directly supporting the combined-effect motivation of this study while also revealing meaningful differences in how individual models weight that combination.

5. Discussion

5.1. Comparison with State-of-the-Art Studies

Direct numerical comparison across cancer classification studies is complicated by differences in cohort, feature modality, and number of target classes, and the comparisons below should be read as contextual positioning rather than a controlled benchmark. With that caveat, the two studies identified in Section 2.3 as the closest prior work in scope — both addressing multi-class cancer-type classification rather than a single binary task — provide the most meaningful reference points. He et al. reported 88.2% accuracy and an F1-score of 88.9% classifying 13 cancer types from a 600-gene somatic-mutation panel across 4,909 samples [7]. Li and Luo reported that their five-classifier weighted ensemble achieved the highest accuracy among the models they compared, classifying 14 cancer types across 6,249 samples using whole-exome sequencing mutation calls [6]. The present study's best-performing model (XGBoost) achieved 87.0% test accuracy (95% CI: 0.850–0.888) and an F1-score of 0.869 across 20 cancer types. This accuracy is within approximately 1.2 percentage points of the figure reported by He et al., despite classifying a substantially larger and more diverse set of cancer types (20 versus 13) and relying on ten routinely available clinical and summary genomic covariates rather than a several-hundred-gene mutation panel or whole-exome sequencing output. This suggests that a smaller, more clinically accessible feature set can support multi-class pan-cancer classification at a scale and accuracy broadly comparable to prior approaches that depend on substantially more extensive genomic sequencing.

Comparison against the single-cancer, binary-classification studies reviewed in Section 2.2 is necessarily more limited: reported accuracies in that body of work range from 83% to 98.3% [13, 19, 20], but these figures reflect two-class problems, for which the chance-level baseline (50%) is far higher than the chance-level baseline for a 20-class problem (5%). A lower raw accuracy relative to these binary studies does not indicate weaker model performance once this difference in task difficulty is accounted for, and no direct claim of superiority over these studies is made here. Rather, the present results are offered as evidence that comparable classification performance to prior single-cancer and 13–14-class pan-cancer studies is achievable at a broader, 20-class scale using a substantially smaller and more routinely collected feature set — a distinction discussed further alongside the study's limitations in Section 5.3.

5.2. Model Comparison and Performance Analysis

Among the four models evaluated, XGBoost achieved the strongest overall performance across nearly all metrics (Section 4.5), suggesting it may be the most promising candidate for further development as a decision-support tool for pan-cancer classification, pending external validation. RF performed closely behind XGBoost, with particularly strong specificity and balanced F1-scores across the 20 cancer types, supporting its suitability for multi-class classification tasks of this kind. Both ensemble methods produced well-calibrated probability outputs, reflected in their comparatively low log loss (Table 6), consistent with their performance on the other metrics reported in Section 4.

SVM performed reliably for some cancer types but showed lower sensitivity and precision overall, particularly for Breast Carcinoma and Gallbladder Carcinoma — classes that, as discussed in Section 5.4, show substantial feature overlap with other cancer types in the dataset. ANN trailed the other three models across most metrics; ANN's performance is especially sensitive to network architecture, activation function choice, and regularization, and tends to require larger training sets than tree-based methods to reach comparable performance, so its relatively weaker results here are consistent with the sample sizes available for several cancer types in this cohort (Section 5.3). Directions for improving on these results — expanded data collection, alternative architectures, and refined feature selection — are discussed in Section 7.

5.3. Limitations

However, despite the positive results achieved by the proposed models, some limitations should be mentioned. First

of all, the study was based solely on one retrospective clinical dataset, which could limit the applicability of the results to other organizations or other patients' groups. Also, considering the fact that the clinical dataset used contains information about Chinese patients, demographic, genetic, and environmental factors could affect the applicability of the proposed models to other ethnicities or geographic regions. Hence, validation on different international populations is necessary before using the model in practice. Second, while the problem of class imbalance was addressed using the ratio-capped oversampling technique (Section 3.1.4), the two smallest classes, Thymic Tumor and Thyroid Carcinoma (each with only 20 real samples in the training fold), remain overwhelmingly synthetic after resampling: the 10× cap yields a target of 200 samples for these classes, far short of the 1,000-sample ceiling applied to larger classes, so real samples make up only about 10% of their final training data. Reported performance on these classes should therefore be interpreted as a reflection of how well the model performs on a training distribution that is mostly synthetic, and the near-perfect scores observed for them (Section 4.7) should be read with this in mind rather than as evidence of strong generalization to unseen real patients from these cancer types. Third, the proposed framework relies solely on structured clinical variables and does not incorporate complementary modalities such as medical imaging or full genomic sequencing, which could further enhance predictive performance. Fourth, DNA Input — a technical sequencing covariate rather than a clinical or biological variable — emerged as one of the most influential predictors in the feature-importance analysis (Section 4.8). While this may reflect genuine biological variation (for example, DNA input correlating with tumor cellularity, which itself varies by cancer type), it may also partly reflect systematic differences in sequencing protocol across cancer types in the source cohort, a possibility that cannot be fully disentangled with the metadata available in this study. Future work will therefore focus on expanding the dataset — particularly collecting additional real samples for the rarest cancer types to reduce reliance on synthetic oversampling — exploring additional network architectures, improving feature selection strategies, investigating whether DNA Input's predictive contribution reflects biological signal or batch-related artifact, and validating the proposed framework using independent multicenter cohorts and prospective clinical studies before routine clinical deployment.

5.4. Class-Wise Performance Patterns

Classification difficulty across cancer types does not track original sample size. Liver Hepatocellular Carcinoma, with approximately 808 real samples in the training fold, requires only modest oversampling under the 10× cap (a target of 1,000, of which roughly 19% is synthetic) and is still among the weakest-performing classes across all four models and metrics. Thymic Tumor and Thyroid Carcinoma, with only 20 real training samples each, are oversampled to a target of 200 — roughly 90% synthetic — yet are among the strongest performers. This indicates that the lower accuracy observed for Kidney Renal Cell Carcinoma, Liver Hepatocellular Carcinoma, Breast Carcinoma, and Soft Tissue Sarcoma reflects genuine overlap in clinical-genomic feature space with other cancer types in the cohort, rather than a limitation of training data volume. It also means the near-perfect scores on the smallest classes warrant caution: with real sample counts this low, and correspondingly few real patients in their held-out test sets, these estimates carry substantial uncertainty regardless of the point value reported.

6. Conclusion

This study evaluated four ML classifiers — RF, XGBoost, SVM, and ANN — for pan-cancer classification across 20 cancer types using a combination of routinely available clinical and summary genomic covariates. XGBoost achieved the strongest overall performance (test accuracy 87.0%, 95% CI: 0.850–0.888; F1-score 0.869), with RF close behind (accuracy 84.7%); both ensemble methods outperformed SVM and ANN by a consistent margin supported by non-overlapping cross-validation confidence intervals at the fold level (Section 4.4). This performance is broadly comparable to prior multi-class pan-cancer classification studies covering a narrower 13–14 cancer types and relying on substantially higher-dimensional genomic inputs such as whole-exome sequencing or large curated mutation panels (Section 5.1), suggesting that a smaller, more clinically accessible set of ten clinical and summary genomic variables can support classification across a broader range of cancer types without requiring extensive sequencing infrastructure.

Rather than assuming that mutation count, diagnosis age, smoking status, tumor purity, tumor stage, and sex jointly contribute to classification performance, this study tested that claim directly through permutation feature importance (Section 4.8), finding that multiple features — including several with only weak pairwise correlation to one another (Section 3.2) — each contribute distinct, non-overlapping predictive value across all four models. This provides empirical support for the combined-effect motivation underlying the study, rather than treating it as an assumption.

At the same time, the results point to specific constraints on how these findings should be interpreted. Performance on the rarest cancer types, such as Thymic Tumor and Thyroid Carcinoma, reflects classification largely on synthetically oversampled training data (Section 5.3), and lower performance on high-support classes such as Liver Hepatocellular Carcinoma and Kidney Renal Cell Carcinoma reflects genuine overlap in clinical-genomic feature space rather than insufficient training data (Section 5.4). The dataset is drawn from a single retrospective Chinese patient cohort, and the models have not yet been evaluated on independent or prospective data. These results should therefore be read as evidence that ML-based pan-cancer classification from routinely available clinical and genomic covariates is feasible and merits further development, not as evidence of readiness for clinical deployment; external validation across diverse, multicenter cohorts remains a necessary next step before these findings can inform clinical decision-making.

7. Future Prospects

Specific directions for future work based on the weaknesses identified in this study can be outlined. First, increasing data gathering efforts for the most rare cancers in the population, specifically Thymic Tumor and Thyroid Carcinoma, will make it less necessary to use artificial oversampling and will allow assessment of the results achieved on a dataset mainly comprised of actual patients' data. Second, the contribution of DNA Input as one of the most influential predictors across models (Section 4.8) warrants direct investigation into whether this reflects genuine biological signal, such as its relationship to tumor cellularity, or a sequencing-protocol artifact specific to this cohort; resolving this would clarify whether DNA Input should be treated as a clinically meaningful predictor or a technical covariate to be controlled for in future work. Third, class imbalance and input dimensionality are separate problems calling for separate remedies: further refinement of the resampling strategy used here (Section 3.1.4), potentially including cost-sensitive learning as an alternative or complement to synthetic oversampling, would directly address class imbalance, while dimensionality reduction and automated feature selection remain a distinct, complementary direction for managing the input space as additional covariates are incorporated in future work.

Beyond the present feature set, incorporating complementary data modalities — medical imaging, full genomic sequencing, or longitudinal treatment and outcome data — represents a natural extension once the present clinical-and-summary-genomic approach has been externally validated, and may further improve classification performance and clinical utility. Any such extension, together with the present framework, would require validation on independent, multi-institutional cohorts spanning diverse geographic and ethnic populations, and ideally prospective clinical studies, before informing real-world diagnostic workflows. Should these steps confirm the feasibility demonstrated here, ML-based pan-cancer classification models of this kind may eventually support clinical decision-making as one input among several — for example, within telemedicine or decision-support platforms — rather than as a standalone diagnostic tool.

References

[1] O. Elemento, C. Leslie, J. Lundin, and G. Tourassi, "Artificial intelligence in cancer research, diagnosis and therapy," Nat. Rev. Cancer, vol. 21, no. 12, pp. 747–752, Dec. 2021, doi: 10.1038/s41568-021-00399-1.

[2] B. Bhinder, C. Gilvary, N. S. Madhukar, and O. Elemento, "Artificial Intelligence in Cancer Research and Precision Medicine," Cancer Discov., vol. 11, no. 4, pp. 900–915, Apr. 2021, doi: 10.1158/2159-8290.CD-21-0090.

[3] Z.-H. Chen, L. Lin, C.-F. Wu, C.-F. Li, R.-H. Xu, and Y. Sun, "Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine," Cancer Commun. Lond. Engl., vol. 41, no. 11, pp. 1100–1115, Nov. 2021, doi: 10.1002/cac2.12215.

[4] M. J. Iqbal et al., "Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future," Cancer Cell Int., vol. 21, no. 1, p. 270, May 2021, doi: 10.1186/s12935-021-01981-1.

[5] Y. Kumar, S. Gupta, R. Singla, and Y.-C. Hu, "A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis," Arch. Comput. Methods Eng. State Art Rev., vol. 29, no. 4, pp. 2043–2070, 2022, doi: 10.1007/s11831-021-09648-w.

[6] Y. Li and Y. Luo, "Performance-weighted-voting model: An ensemble machine learning method for cancer type classification using whole-exome sequencing mutation," Quant. Biol. Beijing China, vol. 8, no. 4, pp. 347–358, Dec. 2020, doi: 10.1007/s40484-020-0226-1.

[7] B. He et al., "A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation," Biochim. Biophys. Acta Mol. Basis Dis., vol. 1866, no. 11, p. 165916, Nov. 2020, doi: 10.1016/j.bbadis.2020.165916.

[8] C. Taoussi, I. Hafidi, and A. Metrane, "Machine Learning and Deep Learning in Health Informatics: Advancements, Applications, and Challenges," vol. 33, no. 5, 2025.

[9] M. M. Srikantamurthy, V. P. S. Rallabandi, D. B. Dudekula, S. Natarajan, and J. Park, "Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning," BMC Med. Imaging, vol. 23, no. 1, p. 19, Jan. 2023, doi: 10.1186/s12880-023-00964-0.

[10] M. Zhao et al., "Bayesian risk prediction model for colorectal cancer mortality through integration of clinicopathologic and genomic data," Npj Precis. Oncol., vol. 7, no. 1, p. 57, Jun. 2023, doi: 10.1038/s41698-023-00406-8.

[11] M. A. Mohammed, A. Lakhan, K. H. Abdulkareem, and B. Garcia-Zapirain, "A hybrid cancer prediction based on multi-omics data and reinforcement learning state action reward state action (SARSA)," Comput. Biol. Med., vol. 154, p. 106617, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106617.

[12] D. P. Kotevski, R. I. Smee, C. M. Vajdic, and M. Field, "Empirical comparison of routinely collected electronic health record data for head and neck cancer-specific survival in machine-learnt prognostic models," Head Neck, vol. 45, no. 2, pp. 365–379, Feb. 2023, doi: 10.1002/hed.27241.

[13] M. Botlagunta et al., "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," Sci. Rep., vol. 13, no. 1, p. 485, Jan. 2023, doi: 10.1038/s41598-023-27548-w.

[14] J. J, H. Su-Cheng, P. Naveen, and E. A. Anaam, "A Comprehensive Review on Cancer Detection and Classification using Medical Images by Machine Learning and Deep Learning Models," JOIV Int. J. Inform. Vis., vol. 8, no. 3–2, pp. 1590–1598, Nov. 2024, doi: 10.62527/joiv.8.3-2.3061.

[15] M. S. Hashim and A. A. Yassin, "Breast Cancer Prediction Using Soft Voting Classifier Based on Machine Learning Models," vol. 50, no. 2, 2023.

[16] M. K. Gould, B. Z. Huang, M. C. Tammemagi, Y. Kinar, and R. Shiff, "Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data," Am. J. Respir. Crit. Care Med., vol. 204, no. 4, pp. 445–453, Aug. 2021, doi: 10.1164/rccm.202007-2791OC.

[17] X. Jiang and C. Xu, "Deep Learning and Machine Learning with Grid Search to Predict Later Occurrence of Breast Cancer Metastasis Using Clinical Data," J. Clin. Med., vol. 11, no. 19, p. 5772, Sep. 2022, doi: 10.3390/jcm11195772.

[18] Q. Chen, D. R. Cherry, V. Nalawade, E. M. Qiao, A. Kumar, A. M. Lowy, D. R. Simpson, and J. D. Murphy, "Clinical Data Prediction Model to Identify Patients with Early-Stage Pancreatic Cancer," JCO Clin. Cancer Inform., vol. 5, pp. 279–287, Mar. 2021, doi: 10.1200/CCI.20.00137.

[19] M. M. Ahamad et al., "Early-Stage Detection of Ovarian Cancer Based on Clinical Data Using Machine Learning Approaches," J. Pers. Med., vol. 12, no. 8, p. 1211, Jul. 2022, doi: 10.3390/jpm12081211.

[20] H. I. Okagbue, P. I. Adamu, P. E. Oguntunde, E. C. M. Obasi, and O. A. Odetunmibi, "Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer," Health Technol., vol. 11, no. 4, pp. 887–893, Jul. 2021, doi: 10.1007/s12553-021-00572-4.

[21] A. Alemu, A. Girma, M. Abebe, and R. Srinivasagan, "Genetic Algorithm-optimized k-nearest Neighbors and Support Vector Machines for Breast Cancer Detection in Resource-constrained Environments," vol. 52, no. 3, 2025.

[22] J. Peng, D. Zou, W. Gong, S. Kang, and L. Han, "Deep neural network classification based on somatic mutations potentially predicts clinical benefit of immune checkpoint blockade in lung adenocarcinoma," Oncoimmunology, vol. 9, no. 1, p. 1734156, 2020, doi: 10.1080/2162402X.2020.1734156.

[23] A. Karlsson et al., "Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit," ESMO Open, vol. 6, no. 3, p. 100175, Jun. 2021, doi: 10.1016/j.esmoop.2021.100175.

[24] N. Dhariwal, R. Hariprasad, and L. M. Sundari, "An Artificial Intelligence Based Approach Toward Predicting Mortality in Head and Neck Cancer Patients with Relation to Smoking and Clinical Data," IEEE Access, vol. 11, pp. 126927–126937, 2023, doi: 10.1109/ACCESS.2023.3331720.

[25] R. O. Alabi, A. A. Mäkitie, M. Pirinen, M. Elmusrati, I. Leivo, and A. Almangush, "Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer," Int. J. Med. Inf., vol. 145, p. 104313, Jan. 2021, doi: 10.1016/j.ijmedinf.2020.104313.

[26] J. Yang et al., "An Integrated Analysis of Tumor Purity of Common Central Nervous System Tumors in Children Based on Machine Learning Methods," Front. Genet., vol. 12, p. 707802, 2021, doi: 10.3389/fgene.2021.707802.

[27] F. Fan, D. Chen, Y. Zhao, H. Wang, H. Sun, and K. Sun, "Rapid preliminary purity evaluation of tumor biopsies using deep learning approach," Comput. Struct. Biotechnol. J., vol. 18, pp. 1746–1753, Jan. 2020, doi: 10.1016/j.csbj.2020.06.007.

[28] L. Hases, A. Ibrahim, X. Chen, Y. Liu, J. Hartman, and C. Williams, "The Importance of Sex in the Discovery of Colorectal Cancer Prognostic Biomarkers," Int. J. Mol. Sci., vol. 22, no. 3, p. 1354, Jan. 2021, doi: 10.3390/ijms22031354.

[29] J. Simon et al., "Role of sex in lung cancer risk prediction based on single low-dose chest computed tomography," Sci. Rep., vol. 13, no. 1, p. 18611, Oct. 2023, doi: 10.1038/s41598-023-45671-6.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[31] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. in Springer Series in Statistics. New York, NY: Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[32] D. T. Larose and C. D. Larose, Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, 2014.

[33] J. R. Quinlan, C4.5: Programs for Machine Learning. Elsevier, 2014.

[34] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[35] A. Liaw and M. Wiener, "Classification and Regression by RandomForest," R News, vol. 2, pp. 18–22, 2002. [Online]. Available: https://cran.r-project.org/doc/Rnews/

[36] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. New York: Chapman and Hall/CRC, 2012. doi: 10.1201/b12207.

[37] D. R. Cutler et al., "Random Forests for Classification in Ecology," Ecology, vol. 88, no. 11, pp. 2783–2792, 2007, doi: 10.1890/07-0539.1.

[38] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[39] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[40] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 67, no. 2, pp. 301–320, 2005.

[41] J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[42] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: Jan. 21, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[43] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[44] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bull. Math. Biophys., vol. 5, no. 4, pp. 115–133, Dec. 1943, doi: 10.1007/BF02478259.

[45] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[46] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html