BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Optimizing unsupervised feature engineering and classification pipelines for differentiated thyroid cancer recurrence prediction

Emmanuel Onah[1*], Uche Jude Eze[2*], Abdullahi Salahudeen Abdulraheem[3], Ugochukwu Gabriel Ezigbo[4], Kosisochi Chinwendu Amorha[5] and Fidele Ntie-Kang[6*]

Abstract

Background Differentiated thyroid cancer (DTC) is a common endocrine malignancy with rising incidence and frequent recurrence, despite a generally favorable prognosis. Accurate recurrence prediction is critical for guiding post-treatment strategies. This study aimed to enhance predictive performance by refining feature engineering and evaluating a diverse ensemble of machine learning models using the UCI DTC dataset.

Methods Unsupervised data engineering—specifically dimensionality reduction and clustering—was used to improve feature quality. Principal Component Analysis (PCA) and Truncated Singular Value Decomposition (t-SVD) were selected based on superior clustering metrics: adjusted Rand Index (ARI > 0.55) and V-measure (> 0.45). These were integrated into classification pipelines using Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Feedforward Neural Network (FNN), and Gradient Boosting (GB). Model performance was evaluated through bootstrapping on an independent test set, stratified 10-fold cross-validation (CV), and subgroup analyses. Metrics included balanced accuracy, F1 score, AUC, sensitivity, specificity, and precision, each reported with 95% confidence intervals (CIs). SHAP analysis supported model interpretability.

Results The PCA-based LR pipeline achieved the best test set performance: balanced accuracy of 0.95 (95% CI: 0.90–0.99), AUC of 0.99 (95% CI: 0.97–1.00), and sensitivity of 0.94 (95% CI: 0.84–1.00). In stratified CV, it maintained strong results (balanced accuracy: 0.86; AUC: 0.97; sensitivity: 0.80), with consistent performance across clinically relevant subgroups. The t-SVD-based LR pipeline showed comparable performance on both test and CV sets. SVM and FNN pipelines also performed robustly (test AUCs > 0.99; CV AUCs > 0.96). RF and KNN had high specificity but slightly lower sensitivity (test: ~0.87; CV: 0.77–0.80). GB pipelines showed the lowest overall performance (test balanced accuracy: 0.86–0.88; CV: 0.85–0.88).

*Correspondence:
Emmanuel Onah
emmanuel.onah.187260@unn.edu.ng; onahemma111@gmail.com
Uche Jude Eze
eze.18@buckeyemail.osu.edu
Fidele Ntie-Kang
fidele.ntie-kang@ubuea.cm

Full list of author information is available at the end of the article

Conclusions Dimensionality reduction via PCA and t-SVD significantly improved model performance, particularly for LR, SVM, FNN, RF and KNN classifiers. The PCA-based LR pipeline showed the best generalizability, supporting its potential integration into clinical decision-support tools for personalized DTC management.

Clinical trial number Not applicable.

Keywords Differentiated thyroid cancer (DTC), Recurrence prediction, Machine learning, Dimensionality reduction, PCA, Logistic regression, Bootstrapping, Cross-validation, SHAP analysis

Introduction

Thyroid cancer is a malignancy that originates in the thyroid gland, a small, butterfly-shaped organ located at the front of the neck, which plays a crucial role in regulating metabolism through hormone production. Although thyroid cancer is relatively rare compared to other cancers, accounting for less than 3% of all new cancer cases in the United States [1], its incidence has been rising globally, likely due to advancements in diagnostic technologies [2]. For instance, in 2020, thyroid cancer was the tenth most commonly diagnosed cancer worldwide, with approximately 570,000 new cases [3]. Women represent about 77% of those diagnosed, demonstrating a notable gender imbalance in the disease's prevalence [4].

The 2022 World Health Organization (WHO) classification of endocrine and neuroendocrine tumors, classified thyroid cancers into three main categories, with the most common being differentiated thyroid cancer (DTC), which includes papillary and follicular thyroid cancers. These types generally have a favorable prognosis and respond well to treatment. Medullary thyroid cancer (MTC), arising from the C-cells of the thyroid, is less common and more challenging to treat. Anaplastic thyroid cancer (ATC) is the rarest and most aggressive form, often diagnosed at an advanced stage with poor prognosis [5]. While thyroid cancer is often treatable, the various subtypes require distinct management approaches, making early detection and accurate classification essential for the best possible outcomes [5].

DTC, as one of the most prevalent endocrine malignancies, has seen a notable increase in incidence over recent decades [2, 6]. While DTC often carries a favorable prognosis when diagnosed early and treated appropriately, recurrence remains a significant clinical concern, affecting 5–30% of patients [7–9]. This recurrence not only complicates follow-up and long-term management but also negatively impacts survival rates and patient quality of life. Consequently, accurately predicting recurrence is vital for tailoring individualized treatment strategies and improving clinical outcomes. Traditionally, recurrence risk in DTC has been evaluated using clinicopathologic parameters, such as tumor size, lymph node involvement, extrathyroidal extension, and metastasis (TNM staging) [10–12]. However, these parameters, while informative, fail to fully capture the complex and multifactorial nature of thyroid cancer, often leading to generalized risk assessments that may not be optimal for every patient [13].

In recent years, the application of machine learning (ML) and deep learning (DL) to oncological data has emerged as a promising approach to overcome these limitations. These methods enable integration of multidimensional datasets—including clinical, demographic, and even genetic information—into predictive frameworks that are capable of uncovering complex patterns in patient outcomes [14]. Recent efforts in this domain, such as the work by Borzooei et al. (2024) [15], evaluated traditional ML classifiers (e.g., Random Forest, SVM, k-NN) using the UCI machine learning DTC dataset. While their study provided valuable benchmarks, it was limited by the use of a narrow feature set (13 clinicopathologic variables) and did not address key challenges such as class imbalance, which is common in recurrence datasets and can lead to biased predictions favoring the majority class [16].

Other recent studies have explored novel data engineering and modeling strategies to improve predictive accuracy and clinical relevance. Clark et al. (2023) [17] demonstrated that integrating SMOTE-based oversampling with ensemble classifiers significantly improved predictive performance in imbalanced thyroid cancer datasets. Furthermore, recent advances in optimization algorithms, such as the application of genetic algorithms and other metaheuristic techniques, including greylag goose optimization and snake optimization are playing a critical role in enhancing model calibration and feature selection, thereby improving predictive performance in clinical applications [18, 19]. For instance, the optimized BPSO model and the hyOPTGB framework have respectively shown significant improvements in predictive accuracy for predicting COVID-19 spread and hepatitis C virus (HCV) detection by integrating hybrid optimization strategies with gradient boosting techniques [20, 21]. Additionally, recent studies have highlighted the growing effectiveness of hybrid models that combine ML and DL approaches with advanced optimization methods to achieve higher diagnostic accuracy in oncology, particularly in thyroid cancer prediction [2, 4]. In this context, interpretable machine learning approaches—such as the use of XGBoost in conjunction with SHAP analysis—have proven valuable for predicting thyroid

cancer recurrence, offering both robust performance and enhanced model transparency for clinical decision-making [22]. These integrative approaches not only boost diagnostic precision but also contribute to the development of more explainable and actionable decision-support systems in healthcare.

Despite these advances, there remains a lack of systematic comparison between linear and nonlinear dimensionality reduction techniques—such as Principal Component Analysis (PCA), Truncated Singular Value Decomposition (t-SVD), Uniform Manifold Approximation and Projection (UMAP), and T-distributed Stochastic Neighbor Embedding (t-SNE)—in the context of thyroid cancer recurrence prediction. Moreover, few studies have jointly evaluated the trade-offs between model interpretability, computational complexity, and predictive performance in a single, integrated pipeline.

In this study, we propose a comprehensive machine learning framework for predicting recurrence in DTC, utilizing an extensive feature set comprising 16 sociodemographic and clinicopathologic variables. To enhance feature representation and reduce noise, we evaluate the effectiveness of various dimensionality reduction techniques—both linear, including Principal Component Analysis (PCA), truncated Singular Value Decomposition (t-SVD), Fast Independent Component Analysis (f-ICA), and Non-Negative Matrix Factorization (NMF), and non-linear, such as Uniform Manifold Approximation and Projection (UMAP), t-distributed Stochastic Neighbor Embedding (t-SNE), Isometric Mapping (Isomap), and Locally Linear Embedding (LLE). Clustering approaches are also incorporated to further refine the feature space. Model performance is assessed using metrics that account for class imbalance, including balanced accuracy, area under the receiver operating characteristic curve (AUC), and F1-score, evaluated through both bootstrapping and stratified 10-fold cross-validation to ensure robust and generalizable results [23]. By systematically comparing these methods, we aim to identify the optimal combinations of feature engineering and predictive modeling techniques, ultimately improving predictive accuracy and contributing to more personalized and effective treatment strategies for thyroid cancer patients.

The novelty of this work lies in: (1) the integrated, side-by-side comparison of linear and non-linear dimensionality reduction strategies. This comparison is critical because, while non-linear methods can capture complex structures, they are often less interpretable and more computationally expensive. We emphasize the balance between interpretability and predictive performance, which is crucial in clinical applications; (2) the inclusion of underexplored sociodemographic features; and (3) the integration of these techniques within classification pipelines tailored to DTC recurrence prediction, offering new insights into optimizing classification pipelines. The implementation of ensemble and interpretable machine learning methods within a clinically relevant framework further strengthens our approach. We anticipate that our findings will contribute to more personalized and accurate risk assessments for patients with thyroid cancer.

Materials and methods

Dataset

The differentiated thyroid cancer dataset was obtained from the UCI Machine Learning Repository [24]. It comprises 383 instances, each characterized by 16 sociodemographic and clinicopathologic features, including age, gender, smoking status, history of smoking, history of radiotherapy, thyroid function, physical examination, adenopathy, pathology, focality, risk, tumor stage (T), node stage (N), metastasis stage (M), overall stage, and treatment response. The data were collected over a period of 15 years, with each patient followed for at least 10 years [24]. The target variable (Recurred) indicates whether or not the cancer has recurred post-treatment. The dataset contains no missing values across any features. However, there is a class imbalance in the target variable, with 108 instances of recurrence and 275 instances of non-recurrence. We selected these 16 sociodemographic and clinicopathologic features based on their availability and relevance to the literature on thyroid cancer recurrence. While these features are widely recognized as predictors of recurrence, we acknowledge that additional features, such as genetic or molecular data, could further improve predictive accuracy [9–11]. A detailed description of each feature in the dataset is provided in the Supplementary File 1.

Feature preprocessing and engineering

The process of converting raw data into mathematical objects that can be understood by machine learning algorithms while retaining the information in the original dataset is referred to as feature extraction or vector construction [25]. As mentioned earlier, our dataset contains 16 sociodemographic and clinicopathologic features, with no missing values, of which only age is a numerical variable, with the rest being categorical variables (see Supplementary File 1 for more details). The numerical variable was normalized using min-max scaling, while the categorical variables were vectorized using one-hot encoding. Prior to predictive model building, we applied various dimensionality reduction techniques, including both linear methods—such as Principal Component Analysis (PCA), Truncated Singular Value Decomposition (t-SVD), Fast Independent Component Analysis (f-ICA), and Non-Negative Matrix Factorization (NMF)—as well as manifold learning techniques—such as T-distributed Stochastic Neighbor Embedding (t-SNE), Isometric

Mapping (Isomap), Uniform Manifold Approximation and Projection (UMAP), and Locally Linear Embedding (LLE) to the preprocessed data. Dimensionality reduction is a well-established technique for mitigating the curse of dimensionality by projecting high-dimensional data onto a lower-dimensional space, thus simplifying models and making them easier to interpret and understand. PCA linearly projects data onto axes capturing maximum variance [26, 27], while t-SVD, a similar method often used for sparse data, factorizes the data matrix. ICA extracts independent components from multivariate signals [28], with f-ICA accelerating computations. NMF seeks non-negative factors, aiding interpretable feature extraction [29]. t-SNE non-linearly maps data, preserving similarities [30], while Isomap focuses on global geometric structure through geodesic distances [31]. UMAP preserves both local and global structure by optimizing a low-dimensional graph representation [32]. LLE maintains local linear relationships by constructing a weighted graph and embedding data accordingly [33]. Initially, the dataset was partitioned into a training set (75%) for hyperparameter tuning and model development, and a holdout test set (25%) for unbiased evaluation. All data engineering steps were applied independently to the training and test sets, ensuring consistency in preprocessing parameters. This rigorous approach ensures that model evaluation on the holdout test set is unbiased and reflective of real-world performance.

Evaluation and selection of dimensionality reduction techniques for classification pipeline development

To select the optimal dimensionality reduction technique for classification pipelines building, we evaluated the quality of clusters in the engineered datasets using three clustering metrics: adjusted Rand Index (ARI), V-Measure, and Silhouette Coefficient using stratified 10-fold CV, which maintains the same proportion of the minority class (recurrence cases) in each fold. ARI assesses the agreement between the clustering results and ground truth labels (U), adjusted for chance, with scores ranging from – 1 (complete disagreement) to 1 (perfect agreement) [34, 35]. V-Measure balances homogeneity (where each cluster contains only members of a single class) and completeness (where all members of a class are assigned to the same cluster), with scores ranging from 0 to 1 (perfect homogeneity and completeness) [36]. Unlike the other two metrics, the Silhouette coefficient evaluates intrinsic cluster quality without relying on ground truth, measuring how similar data points are to their own cluster compared to others, with scores ranging from – 1 to 1 (higher values indicate better-defined clusters) [37]. K-means clustering was applied to each dataset to generate predicted clusters (V), enabling the calculation of ARI and V-Measure. Cluster centroids and the variance in the first principal components (PC1) of each engineered feature were also analyzed to aid in selecting the optimal dimensionality reduction technique to use in the classification models. The engineered features with the best performance across these metrics were chosen for the classification pipelines building. The mathematical formulations of these metrics are provided below.

$$ARI = \frac{RI \; - \; E\left[RI\right]}{\left(RI\right) \; - \; E\left[RI\right]},$$

Where $\boldsymbol{RI}$ is the Rand Index, calculated as the proportion of pairs of points that are either in the same cluster in both the ground truth (U) and the predicted clustering (V), or in different clusters in both and $\boldsymbol{E}\left[\boldsymbol{RI}\right]$ is the expected value of the Rand Index for random clustering.

$$V - Measure = 2 \times \frac{h \; \times \; c}{h \; + \; c},$$

Where:

$$h \;\left(homogeneity\right) = 1 - \frac{H\left(K\right)}{H\left(C\right)},$$

with $H\left(C|\,K\right)$ being the conditional entropy of the classes given the clusters, and $H\left(C\right)$ being the entropy of the classes, and

$$c \;\left(completeness\right) = 1 - \frac{H\left(C\right)}{H\left(K\right)},$$

with $H\left(K|\,C\right)$ being the conditional entropy of the clusters given the classes, and $H\left(K\right)$ being the entropy of the clusters.

$$Silhouette\; coefficient\;\left(i\right) = \frac{b\left(i\right) \; - \; a\left(i\right)}{\left(a\left(i\right), \;\; b\left(i\right)\right)},$$

Where $a\left(i\right)$ is the average distance between i and all other points in the same cluster and $b\left(i\right)$ is the average distance between i and all points in the nearest neighboring cluster. The overall Silhouette Score is the mean of $s\left(i\right)$ for all data points i.

Classification pipelines, hyperparameter tuning and evaluation

We employed six classifiers, namely Logistic Regression (LR), Gradient Boosting (GB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and a Feedforward Neural Network (FNN) using a pipeline with each of the best performing dimensionality reduction techniques explored. These classifiers were chosen for their diverse approaches to classification,

which range from linear methods to ensemble learning and neural networks, providing a comprehensive comparison of different algorithms for the task at hand. LR, valued for its interpretability by linking classifier coefficients to log-odds [38], struggles with non-linear data. GB, an ensemble method that builds classifiers sequentially to correct errors, excels in handling complex patterns, particularly in imbalanced datasets [39, 40]. SVM constructs hyperplanes in high-dimensional spaces, offering robustness against overfitting and adaptability through kernel functions [41]. RF, another ensemble technique, creates multiple decision trees to reduce overfitting while effectively managing high-dimensional data [42]. KNN, though straightforward in classifying samples based on the majority class of nearest neighbors, can be computationally intensive and sensitive to the choice of k [43]. Lastly, FNN, capable of modeling complex non-linear relationships through layered neurons, are well-suited for diverse tasks but demand careful hyperparameter tuning and significant computational resources [44].

To optimize the performance of each classifier, we conducted an exhaustive search over the hyperparameters. Initially, a wide range of hyperparameters was explored to identify the best combination that maximizes classifier performance. Subsequently, fine-tuning was performed by narrowing the search to specific intervals around the initially identified best parameters. Grid search was implemented with stratified 10-fold CV to ensure that the results were robust and generalizable [45]. CV is a critical step to avoid overfitting and to ensure that the classifier performs well on unseen data [46].

The classification pipelines' performance was first evaluated on a holdout test set, reserved from the original dataset, to ensure that the pipelines were assessed on data they had not seen during training. Given the imbalanced nature of the dataset, we employed a comprehensive set of evaluation metrics that are sensitive to class imbalance. These included balanced accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, specificity, precision, and F1 score. Confusion matrices for the various classification pipelines were also generated to provide a granular view of classification performance, particularly in terms of correctly and incorrectly predicted instances across the positive and negative classes. Balanced accuracy was particularly emphasized as it accounts for imbalances by averaging the recall obtained on each class [47, 48]. Additionally, AUC was used to assess the trade-off between true positive and false positive rates across different threshold settings [49, 50]. Sensitivity (recall) and specificity provided insights into the pipelines' ability to correctly identify positive and negative cases, respectively [25, 51, 52], while precision and F1-score were used to evaluate the relevance of the positive predictions [52]. Additionally, the Detection Error Tradeoff (DET) curve, which is a plot that shows the trade-off between false positive rate (FPR) and false negative rate (FNR) for a binary classifier was employed to further evaluate the classification pipelines' performance. To enhance the statistical robustness of our classification pipelines evaluation, we computed 95% confidence intervals (CIs) for the performance metrics. Confidence intervals were estimated using a bootstrap resampling technique with 1,000 iterations, where the test dataset was resampled with replacement in each iteration. This non-parametric approach provides a robust estimation of the variability in performance and quantifies the uncertainty around each metric, supporting more reliable comparisons and interpretations [53]. The resulting CIs are reported alongside the point estimates in the results section.

To further ensure the robustness and generalizability of our classification pipelines, stratified 10-fold CV was employed on the entire dataset. Stratification was particularly important in this context to maintain the distribution of classes across all folds, ensuring that each fold was representative of the overall class distribution. This method provided a more reliable estimate of the classification pipelines' performance, especially in scenarios with imbalanced data, and helped identify models that generalize well beyond the specific dataset used [47, 48].

We employed a class weighting technique to address the class imbalance in our dataset during the classification pipelines' training. Specifically, we set the class weight attribute to "balanced," allowing the pipelines to place greater emphasis on the minority class, thereby mitigating bias toward the majority class. For KNN, which does not support class weighting, we set the weight parameter to "distance," giving closer neighbors more influence, which can improve performance on imbalanced datasets. Similarly, early stopping was applied in the FNN to prevent it from overfitting to the majority class. In addition, we conducted rigorous hyperparameter tuning and used a range of performance metrics, including balanced accuracy, AUC, sensitivity, specificity, precision, and F1 score (all of which are sensitive to class imbalance). This ensured a robust evaluation of our classification pipelines' effectiveness in handling class imbalance. These metrics are defined by the following formulas, where the symbols and notations have their usual meaning.

$$Balanced\,accuracy = \frac{1}{2}\left(Sensitivity + Specificity\right),$$

$$AUC = \int_0^1 TPR\left(FPR\right) dFPR$$

$$Sensitivity = \frac{TP}{TP + FN},$$

$$Specificity = \frac{TN}{TN + FP},$$

$$Precision = \frac{TP}{TP + FP},$$

$$F1\ Score = 2\left(\frac{Precision \times Sensitivity}{Precision + Sensitivity}\right)$$

Stratified subgroup evaluation of the best-performing classification pipeline

Following the construction of the classification pipelines using the best-performing dimensionality reduction techniques, a pairwise comparison of the pipelines was conducted using the Wilcoxon signed-rank test to determine if there were any statistically significant differences ($p < 0.05$) in their predictive performance, with balanced accuracy used as the representative metric for comparison. This non-parametric test is particularly well-suited for comparing paired, non-normally distributed data, which is often the case with machine learning performance metrics [54]. To assess the robustness and clinical applicability of the best-performing classification pipeline, we conducted stratified validation across several clinically meaningful subgroups. These included age cohorts (<45, 45–60, >60), ATA guideline-based risk categories (Low, Intermediate, High), TNM staging components (T, N, and M classifications), adenopathy groups (No, Right, Bilateral, Left, Extensive, or Posterior), pathology categories (Papillary, Micropapillary, Follicular, or Hurthel cell), and focality groups (Uni-Focal or Multi-Focal). Subgroups were defined based on the available clinical metadata in the dataset. For each subgroup, we computed a suite of evaluation metrics—balanced accuracy, F1 score, ROC AUC, sensitivity, specificity, and precision—using the trained classification pipeline on the corresponding test data partition. Subgroups with insufficient class variation (i.e., only one class present in the outcome variable) were excluded from metric computation to ensure validity.

Best-performing classification pipeline explainability using SHAP analysis

To enhance the interpretability and clinical trustworthiness of our best-performing classification pipeline, we conducted a post hoc explainability analysis using SHAP (SHapley Additive exPlanations). SHAP provides a unified, game-theoretic framework for explaining the output of machine learning models by quantifying the marginal contribution of each feature to individual predictions [22, 55]. Since the principal components used in our dimensionality reduction approach are abstract transformations lacking direct clinical meaning, applying SHAP directly to these components would offer limited interpretability. To address this, we applied SHAP to the encoded and preprocessed clinical features with the dimensionality reduction step excluded. This allowed for the attribution of model predictions to original, clinically relevant variables such as age, thyroid function, focality, and adenopathy. SHAP values were computed for each instance and summarized to provide a global feature importance plot. This methodology supports model transparency, facilitates clinical interpretation, and aligns with emerging standards for explainable AI in healthcare. Figure 1 illustrates the workflow of the feature engineering techniques and classification modeling methodologies used in the study.

Software and computational tools

The codebase was implemented entirely in Python 3.8.10 [56]. The dimensionality reduction technique, clustering, modeling, and evaluation were implemented using scikit-learn version 1.2.2 [57], SciPy version 1.7.3 [58], NumPy version 1.21.2 [59], and UMAP version 0.5.1 [32]. Data manipulations and visualization were performed with Pandas version 1.3.3 [60], Matplotlib version 3.4.3 [61], Seaborn version 0.11.2 [62], while model explanation was carried out using SHAP version 0.47.1 [55].

Results and discussion

In this study, we employed a variety of unsupervised machine learning techniques, including dimensionality reduction methods such as PCA, f-ICA, t-SVD, NMF, UMAP, t-SNE, Isomap, LLE, and K-means clustering, to engineer features for classification pipelines designed to predict thyroid cancer recurrence in patients' post-treatment. The dataset used was the differentiated thyroid cancer dataset from the UCI Machine Learning Repository [24]. We selected a diverse set of classifiers including LR, GB, SVM, RF, KNN, and FNN due to their different learning paradigms. We have emphasized that stratification was employed to ensure that the distribution of recurrence and non-recurrence examples remains uniform across all 10 folds during the stratified 10-fold CV trials. Additionally, evaluating multiple cross-validated performance metrics is recognized as a best practice, as it offers a more comprehensive and unbiased assessment of model performance, mitigating the risk of relying on a single metric that may be skewed by a particular subset of the data [25]. This method reduces the likelihood of overfitting. To further ensure the robustness of our findings, the performance of the classification pipelines was also assessed on an independent test set that had not been previously exposed to the models, providing a more accurate measure of predictive

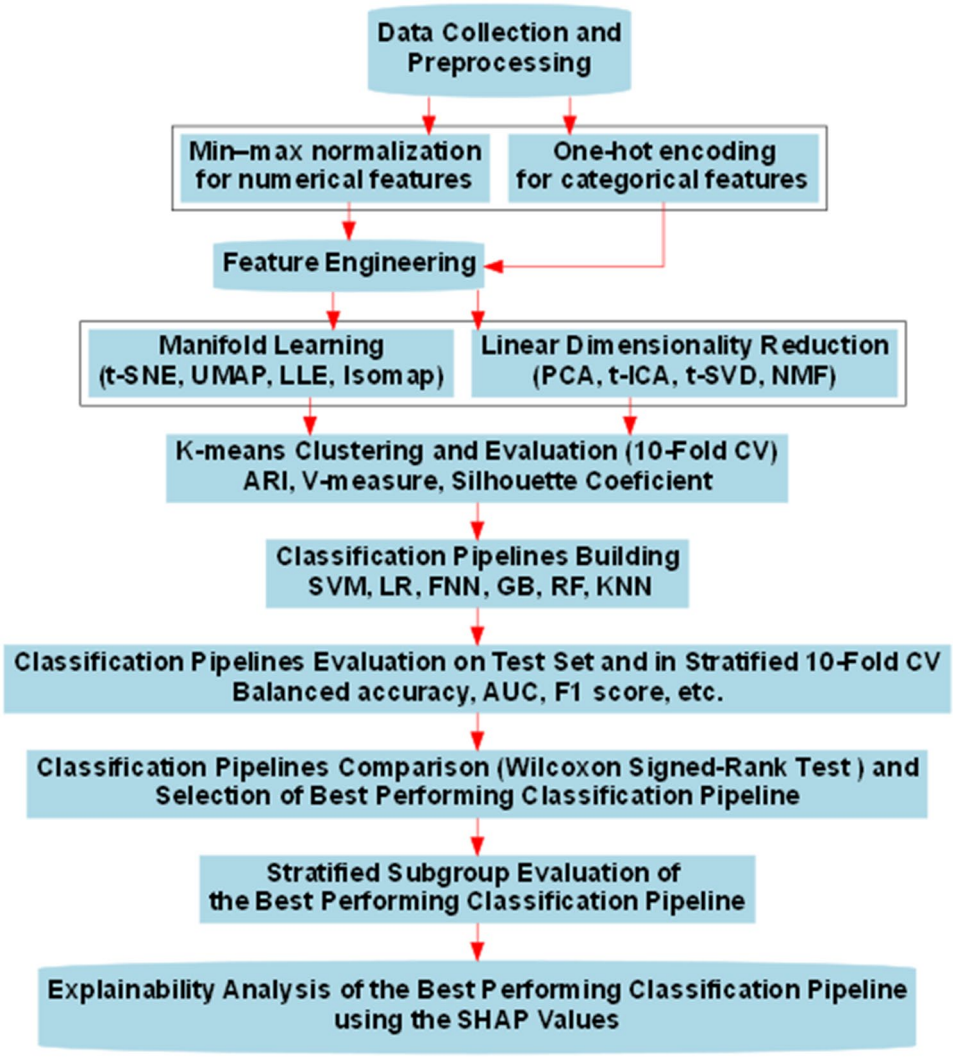

Fig. 1 Workflow of feature engineering techniques and classification modeling methodologies employed in the study. PCA = Principal Component Analysis, t-SVD = Truncated Singular Value Decomposition, f-ICA = Fast Independent Component Analysis, NNM = Non-Negative Matrix Factorization, t-SNE = T-distributed Stochastic Neighbor Embedding, Isomap = Isometric Mapping, UMAP = Uniform Manifold Approximation and Projection, LLE = Locally Linear Embedding (LLE), ARI = Adjusted Rand Index, LR = Logistic Regression, GB = Gradient Boosting, SVM = Support Vector Machine, RF = Random Forest, KNN = K-Nearest Neighbors, FNN = Feedforward Neural Network, AUC = area under the receiver operating characteristic (ROC) curve, SHAP = SHapley Additive exPlanations

strength. The performance metrics reported are mean values accompanied by their 95% CIs, calculated across the folds in the stratified 10-fold CV loop and through bootstrap resampling for the test set. Although computationally intensive, this approach minimizes data wastage and enhances the reliability of the estimates [63]. We conducted stratified validation across several clinically meaningful subgroups—including age, ATA guideline-based risk categories, TNM staging, presence of adenopathy, pathological subtype, and tumor focality—to assess the consistency and robustness of the best-performing classification pipeline across diverse patient populations. To further enhance interpretability, we employed SHAP analysis to quantify the contribution of individual features to classification pipeline predictions. This approach provided clear insights into how specific clinical and pathological variables influenced recurrence risk, thereby reinforcing the clinical relevance and potential utility of the proposed pipeline as a decision-support tool in personalized DTC management.

Feature engineering using dimensionality reduction

To identify the optimal dimensionality reduction technique for building the classification pipelines, we established a selection criterion based on the ARI and V-measure scores, using an arbitrary cut-off of 0.4 for both metrics. This threshold was set to ensure that the clusters resulting from each dimensionality reduction

technique reflected meaningful patterns in the dataset rather than artifacts introduced by the technique itself. K-means clustering (with k = 2) was applied to identify cluster centroids and assign labels to data points. ARI and V-measure were selected over the Silhouette coefficient for this evaluation because they compare clustering results against ground truth labels. Specifically, ARI measures the degree of agreement between the predicted clusters and true labels, adjusting for chance [34, 35], while V-measure assesses the trade-off between homogeneity (each cluster contains only members of a single class) and completeness (all members of a class are assigned to the same cluster) [36]. In contrast, the Silhouette coefficient evaluates intrinsic cluster quality without considering ground truth, by measuring how similar a point is to its own cluster compared to others [37]. Therefore, ARI and V-measure were deemed more appropriate for our goal of aligning clustering structure with known class labels.

Table 1 provides the scores for each method across the clustering metrics utilized. Among the methods evaluated, PCA and t-SVD exhibited the highest performance across all clustering metrics, both achieving ARI scores of 0.557 and 0.558, respectively, and V-measure scores of 0.451 and 0.459. Notably, t-SVD's silhouette coefficient of 0.537 was slightly higher than PCA's 0.489, underscoring its potential for capturing meaningful clusters. Techniques such as t-SNE and Isomap followed, with ARI, V-measure, and silhouette coefficients ranging between 0.258 and 0.292, 0.277–0.292, and 0.334–0.362, respectively. These methods demonstrated moderate clustering performance, albeit with higher variance in the first principal component (PC1) (see Table 1). The poor performance of f-ICA and NMF, with ARI and V-measure scores below 0.2 and silhouette coefficients below 0.36, indicated that these techniques were less effective at capturing the intrinsic structure of the dataset.

Table 1 Performance metrics of the engineered data clustering

Method	ARI	V-Measure	Silhouette coefficient	PC1 variance
PCA*	0.557	0.451	0.489	1.200
f-ICA	0.179	0.165	0.318	1.001
t-SVD*	0.558	0.459	0.537	0.537
NMF	0.013	0.102	0.352	0.156
UMAP	-0.076	0.093	0.604	2.565
t-SNE	0.258	0.277	0.362	22.727
Isomap	0.258	0.292	0.334	4.477
LLE	-0.081	0.083	0.633	0.049

The methods marked with asterisks (*) are the ones adopted in the classification pipelines. The choice is based on the performance of the clustering metrics, PC1 variance, and how distinct the clusters are (see Fig. 2). PCA = Principal Component Analysis, t-SVD = Truncated Singular Value Decomposition, f-ICA = Fast Independent Component Analysis, NNM = Non-Negative Matrix Factorization, t-SNE = T-distributed Stochastic Neighbor Embedding, Isomap = Isometric Mapping, UMAP = Uniform Manifold Approximation and Projection, LLE = Locally Linear Embedding, PC1 = first principal components

Interestingly, UMAP and LLE, despite their negative ARI scores and low V-measure values, achieved the highest silhouette scores, exceeding 0.6. This suggests that embedding methods such as UMAP and LLE may better preserve the intrinsic properties of the data when compared to linear techniques like PCA and t-SVD. However, their negative ARI scores indicate that the clusters identified by these methods diverged significantly from the ground truth labels. Also, while these non-linear dimensionality reduction techniques are excellent for non-linear data, they are computationally intensive and difficult to interpret in clinical settings [30–33]. PCA and t-SVD, on the other hand, offer a good balance between retaining important variance and being interpretable and efficient in high-dimensional datasets [26–28]. Moreover, our analysis shows that PCA and t-SVD outperformed non-linear techniques in clustering metrics such as ARI and V-Measure for this specific dataset.

Furthermore, the variance of the PC1 across the various methods (Table 2) revealed that PCA and t-SVD, which clustered data points most closely with respect to the ground truth, exhibited low to moderate variance. In contrast, methods like t-SNE and Isomap showed significantly higher variance, which could reflect their ability to capture more complex, non-linear patterns in the data. As depicted in Figs. 2 and 3, the clustering distinctiveness and the spread of PC1 across these methods provide further insight into their respective capabilities. Based on these findings, PCA and t-SVD were selected as the optimal dimensionality reduction techniques for the classification pipelines. The superior clustering performance of these methods, combined with their moderate PC1 variance, suggests they strike a balance between capturing meaningful data patterns and maintaining the interpretability of the resulting features.

Distributions of the engineered features

The distribution of the first principal component (PC1) derived from various dimensionality reduction techniques, as illustrated in Fig. 3, Panel A, exhibits a distinct bimodal pattern. This bimodality is significant as it suggests the presence of two latent subpopulations within the dataset. The characteristics of these distributions—such as their height, sharpness (kurtosis), and width—provide insights into the underlying data structure and the effectiveness of each dimensionality reduction method.

Techniques like PCA (curve 'a') and t-SVD (curve 'c'), which clustered data points most closely with respect to the ground truth, display moderate to sharp peaks in their bimodal distributions, indicating that the variance captured by these methods is concentrated around two distinct clusters with minimal overlap. The sharpness of these peaks suggests that these methods decompose

Table 2 Optimal hyperparameters discovered for each classification pipeline

Classifier	Hyperparameter	PCA-based classification pipeline	t-SVD-based classification pipeline
RF	criterion	entropy	entropy
	max_depth	None	None
	min_sample_leaf	4	3
	class_weight	balanced	balanced
	sample_split	4	6
	n_estimators	403	400
	max_features	log2	log2
GB	criterion	friedman_mse	squared_error
	learning_rate	0.35	0.3
	loss	exponential	log_loss
	class_weight	balanced	balanced
	max_depth	6	4
	n_estimators	150	125
SVM	C	0.1	0.25
	class_weight	balanced	balanced
	kernel	sigmoid	sigmoid
LR	C	0.35	0.1
	solver	liblinear	liblinear
	class_weight	balanced	balanced
	max_iter	5000	5000
	penalty	l1	l2
KNN	n_neighbors	17	18
	weight	distance	distance
	p	2	4
FNN	alpha	0.5	1.0
	activation	relu	identity
	hidden_layer_size	(300,)	(125,155)
	learning_rate	adaptive	constant
	early_stopping	True	True
	solver	lbfgs	adam
	max_iter	10,000	15,000

LR = Logistic Regression, GB = Gradient Boosting, SVM = Support Vector Machine, RF = Random Forest, KNN = K-Nearest Neighbors, FNN = Feedforward Neural Network

features that are highly informative, leading to a clear separation between the two underlying groups. This clear separation is crucial in a binary classification context, as it enhances the discriminative power of the classifier by providing a strong signal corresponding to each class. The concentrated variance around the two modes reinforces the idea that PCA and t-SVD effectively capture the intrinsic structure of the data, making them suitable for feature engineering in this context.

In contrast, techniques like t-SNE (curve 'f') and Isomap (curve 'g') produce broader, lower peaks. This broader distribution implies a more gradual separation between the two clusters, with a higher degree of overlap. The lower height of the peaks suggests that these methods capture a more diffuse variance, possibly reflecting non-linear relationships in the data that are less sharply defined. While capturing these complex patterns can be valuable, it may also indicate that these methods are less effective in creating a clear-cut separation between the classes. This could introduce ambiguity in the classification task, potentially leading to reduced performance of the classification pipelines.

The heatmap in Fig. 3, Panel B complements this distribution analysis by visually representing how the standardized PC1 values vary across the dataset. Techniques like PCA, t-SVD, and NMF exhibit abrupt transitions between high and low PC1 values, consistent with the sharp peaks observed in the density plots. This abruptness reflects the strong underlying structure captured by these techniques, clearly distinguishing the two subpopulations. On the other hand, methods such as Isomap and LLE show more gradual transitions in the heatmap, with a smoother gradient of PC1 values. This corresponds to the broader peaks in the density plots and suggests a more nuanced capture of the data's structure, potentially blending the two subpopulations together more than the other techniques.

These varying characteristics of the PC1 distributions and heatmap patterns across different dimensionality reduction techniques underscore the importance of technique selection in the feature engineering process. Techniques that produce sharp, well-separated bimodal distributions, such as PCA and t-SVD, are likely to yield features that are more effective for binary classification tasks due to their ability to create a clearer distinction between classes. Conversely, methods that produce broader distributions, such as Isomap and LLE, may capture more complex non-linear relationships but could introduce greater ambiguity in class separation, potentially impacting the performance of the classification pipelines.

While the observed bimodality in the distributions is encouraging, it is essential to recognize potential limitations. The robustness of these findings requires further investigation, particularly in elucidating the underlying biological mechanisms driving the observed heterogeneity. Future research should explore additional clustering algorithms and incorporate external data sources to validate and refine the identified subgroups. A deeper understanding of these subpopulations could lead to the development of more precise classification pipelines, ultimately improving patient outcomes.

Hyperparameter optimization of the classification pipelines

Following the identification of PCA and t-SVD as the optimal dimensionality reduction techniques, classification pipelines were constructed using each technique for feature extraction. For both PCA and t-SVD, only the first three principal components were selected as input

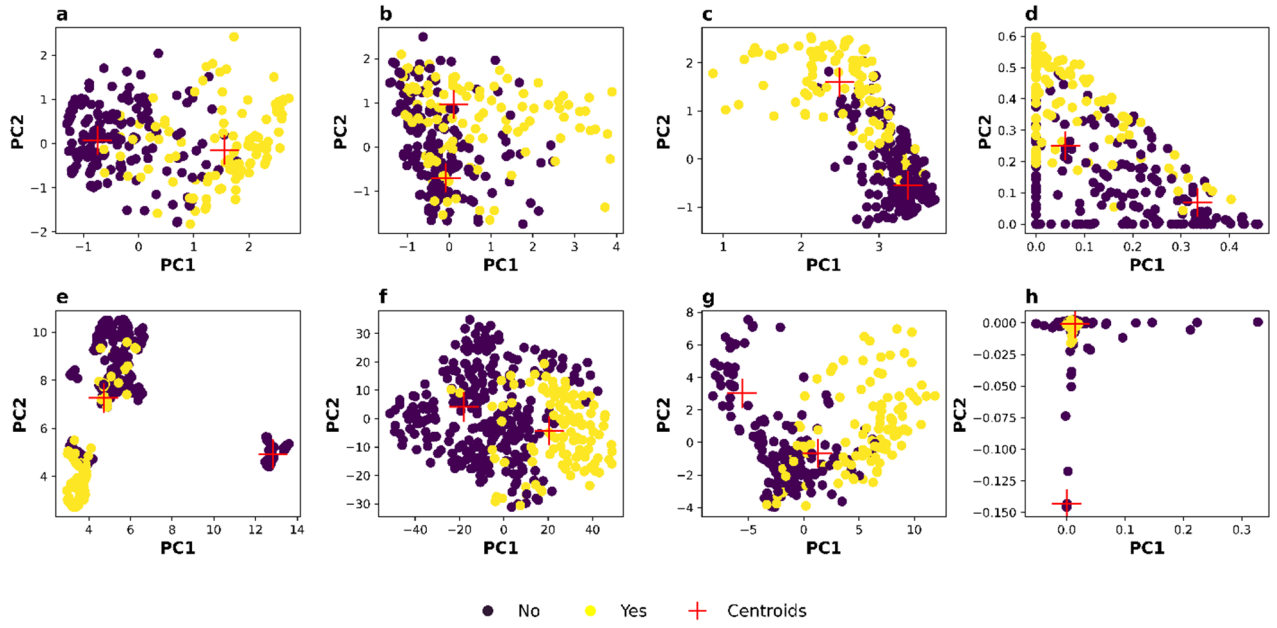

Fig. 2 Scatter plots illustrating the dataset clusters from each dimensionality reduction technique. (**a**) PCA-Decomposed data, (**b**) f-ICA-Decomposed data, (**c**) t-SVD-Decomposed data, (**d**) NMF-Decomposed data (**e**) UMAP-Decomposed data (**f**) t-NSE-Decomposed data (**g**) Isomap-Decomposed data (**h**) LLE-Decomposed data. The red "**+**" indicates K-means predicted cluster centers. The data points coloring is in accordance with the grand truth (class labels in the dataset) not the K-means predicted classes

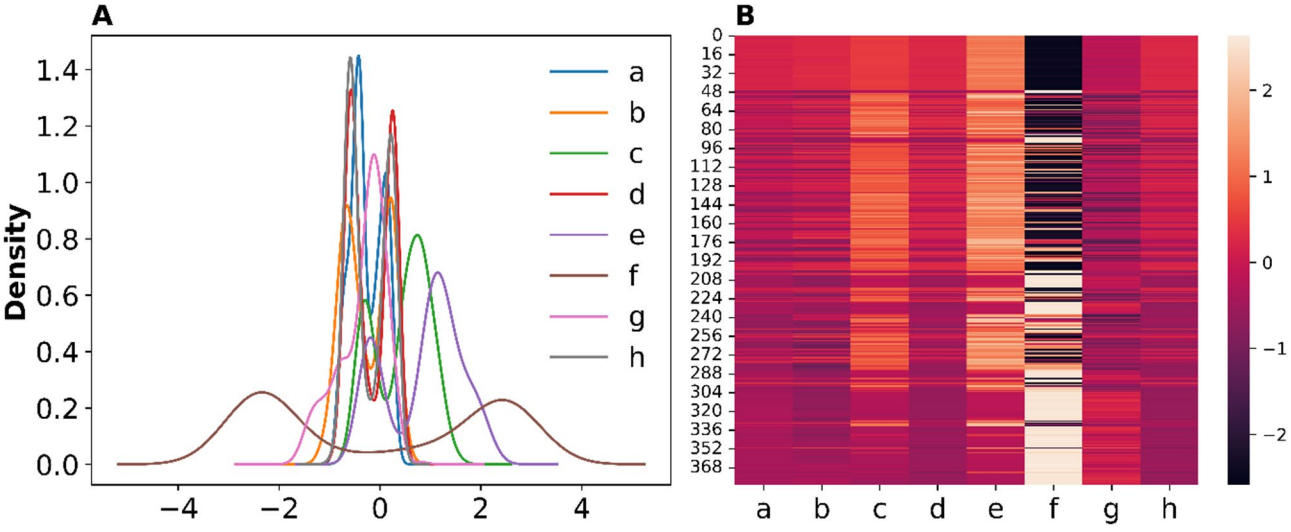

Fig. 3 (**A**) Density distributions and (**B**) heatmap of the standardized first principal components (PC1) derived from features obtained through various dimensionality reduction techniques. The subplots correspond to: (**a**) PC1 from PCA-derived features, (**b**) PC1 from f-ICA-derived features, (**c**) PC1 from t-SVD-derived features, (**d**) PC1 from NMF-derived features, (**e**) PC1 from UMAP-derived features, (**f**) PC1 from t-SNE-derived features, (**g**) PC1 from Isomap-derived features, and (**h**) PC1 from LLE-derived features. PC1 = First Principal Component

features for the classification pipelines. Before model development, an exhaustive grid search with 10-fold CV was performed to optimize the hyperparameters of each classification pipeline. Initially, broader ranges for each hyperparameter were explored to ensure comprehensive coverage. Subsequently, fine-tuning was carried out with more granular adjustments around the most promising hyperparameter values. Table 2 presents the optimal hyperparameters that resulted in the best performance on both the test set and during 10-fold CV.

Classification pipelines evaluation

The evaluation of the PCA- and t-SVD-based classification pipelines provides valuable insights into their

Table 3 Performance of the PCA- and t-SVD-based classification pipelines on the test set

Classifier	B. Acc. (95% CI)	F1 score (95% CI)	AUC (95% CI)	Sen. (95% CI)	Spec. (95% CI)	Prec. (95% CI)
PCA-based classification pipelines						
RF	0.905 (0.828–0.968)	0.926 (0.870–0.979)	0.977 (0.951–0.996)	0.841 (0.692–0.964)	0.969 (0.922–1.000)	0.929 (0.874–0.979)
GB	0.865 (0.786–0.938)	0.885 (0.820–0.947)	0.956 (0.910–0.990)	0.807 (0.657–0.931)	0.923 (0.848–0.983)	0.887 (0.822–0.947)
SVM	0.936 (0.875–0.985)	0.937 (0.886–0.979)	0.991 (0.975–1.000)	0.936 (0.839–1.000)	0.937 (0.877–0.985)	0.940 (0.890–0.981)
LR	0.953 (0.897–0.993)	0.958 (0.916–0.990)	0.991 (0.974–1.000)	0.936 (0.839–1.000)	0.969 (0.925–1.000)	0.959 (0.916–0.990)
KNN	0.913 (0.844–0.970)	0.927 (0.873–0.979)	0.986 (0.964–0.999)	0.872 (0.735–0.973)	0.954 (0.895–1.000)	0.928 (0.874–0.979)
FNN	0.929 (0.868–0.977)	0.928 (0.876–0.969)	0.981 (0.956–0.997)	0.934 (0.833–1.000)	0.924 (0.859–0.984)	0.932 (0.882–0.972)
t-SVD-based classification pipelines						
RF	0.920 (0.852–0.976)	0.937 (0.883–0.979)	0.982 (0.957–0.996)	0.871 (0.741–0.969)	0.969 (0.922–1.000)	0.938 (0.887–0.980)
GB	0.882 (0.799–0.950)	0.897 (0.832–0.958)	0.970 (0.934–0.993)	0.839 (0.694–0.962)	0.924 (0.859–0.984)	0.899 (0.837–0.958)
SVM	0.928 (0.859–0.983)	0.947 (0.895–0.990)	0.992 (0.976–1.000)	0.872 (0.733–0.971)	0.984 (0.952–1.000)	0.950 (0.905–0.990)
LR	0.944 (0.880–0.992)	0.958 (0.914–0.990)	0.993 (0.979–1.000)	0.904 (0.781–1.000)	0.984 (0.952–1.000)	0.959 (0.917–0.990)
KNN	0.913 (0.844–0.970)	0.927 (0.873–0.979)	0.988 (0.970–0.999)	0.872 (0.735–0.973)	0.954 (0.895–1.000)	0.928 (0.874–0.979)
FNN	0.822 (0.741–0.902)	0.876 (0.803–0.946)	0.993 (0.979–1.000)	0.644 (0.481–0.805)	1.000 (1.000–1.000)	0.903 (0.861–0.951)

B. Acc.=Balanced Accuracy, AUC=area under the receiver operating characteristic (ROC) curve, Sen.=Sensitivity, Spec.=Specificity, Prec.=Precision, LR=Logistic Regression, GB=Gradient Boosting, SVM=Support Vector Machine, RF=Random Forest, KNN=K-Nearest Neighbors, FNN=Feedforward Neural Network, CI=Confidence Interval

Table 4 Performance of the PCA- and t-SVD-based classification pipelines in the stratified 10-fold CV

Classifier	B. Acc. (95% CI)	F1 score (95% CI)	AUC (95% CI)	Sen. (95% CI)	Spec. (95% CI)	Prec. (95% CI)
PCA-based classification pipelines						
RF	0.867 (0.826–0.908)	0.882 (0.843–0.921)	0.965 (0.948–0.981)	0.798 (0.712–0.885)	0.957 (0.935–0.979)	0.864 (0.802–0.926)
GB	0.859 (0.824–0.894)	0.861 (0.827–0.896)	0.953 (0.936–0.970)	0.797 (0.737–0.857)	0.917 (0.881–0.953)	0.808 (0.735–0.882)
SVM	0.873 (0.828–0.918)	0.860 (0.821–0.900)	0.961 (0.940–0.981)	0.845 (0.761–0.928)	0.901 (0.881–0.922)	0.770 (0.725–0.816)
LR	0.859 (0.819–0.899)	0.856 (0.821–0.891)	0.965 (0.950–0.980)	0.798 (0.716–0.880)	0.920 (0.900–0.940)	0.800 (0.763–0.838)
KNN	0.865 (0.827–0.903)	0.869 (0.836–0.903)	0.959 (0.940–0.979)	0.788 (0.709–0.867)	0.942 (0.927–0.957)	0.843 (0.812–0.875)
FNN	0.822 (0.741–0.902)	0.876 (0.803–0.946)	0.993 (0.979–1.000)	0.644 (0.481–0.805)	1.000 (1.000–1.000)	0.903 (0.861–0.951)
t-SVD-based classification pipelines						
RF	0.877 (0.861–0.951)	0.879 (0.861–0.951)	0.967 (0.861–0.951)	0.788 (0.861–0.951)	0.953 (0.861–0.951)	0.877 (0.861–0.951)
GB	0.882 (0.861–0.951)	0.880 (0.861–0.951)	0.958 (0.861–0.951)	0.835 (0.861–0.951)	0.935 (0.861–0.951)	0.852 (0.861–0.951)
SVM	0.848 (0.861–0.951)	0.843 (0.861–0.951)	0.961 (0.861–0.951)	0.780 (0.861–0.951)	0.916 (0.861–0.951)	0.783 (0.861–0.951)
LR	0.854 (0.861–0.951)	0.859 (0.861–0.951)	0.965 (0.861–0.951)	0.770 (0.861–0.951)	0.938 (0.861–0.951)	0.832 (0.861–0.951)
KNN	0.860 (0.861–0.951)	0.868 (0.861–0.951)	0.952 (0.861–0.951)	0.770 (0.861–0.951)	0.949 (0.861–0.951)	0.862 (0.861–0.951)
FNN	0.805 (0.861–0.951)	0.854 (0.861–0.951)	0.962 (0.861–0.951)	0.677 (0.861–0.951)	0.964 (0.861–0.951)	0.913 (0.861–0.951)

B. Acc.=Balanced Accuracy, AUC=area under the receiver operating characteristic (ROC) curve, Sen.=Sensitivity, Spec.=Specificity, Prec.=Precision, LR=Logistic Regression, GB=Gradient Boosting, SVM=Support Vector Machine, RF=Random Forest, KNN=K-Nearest Neighbors, FNN=Feedforward Neural Network, CI=Confidence Interval

effectiveness in predicting thyroid cancer recurrence. Both dimensionality reduction techniques enhanced the classification pipelines performance, though with differing implications for clinical applications. Variations across classification pipelines and performance metrics highlight the relative strengths of each classifier and reduction method. Tables 3 and 4 summarize the performance of the PCA- and t-SVD-based pipelines, across six evaluation metrics for the test set and stratified 10-fold CV, respectively. Figure 4 (panels A–D) presents the

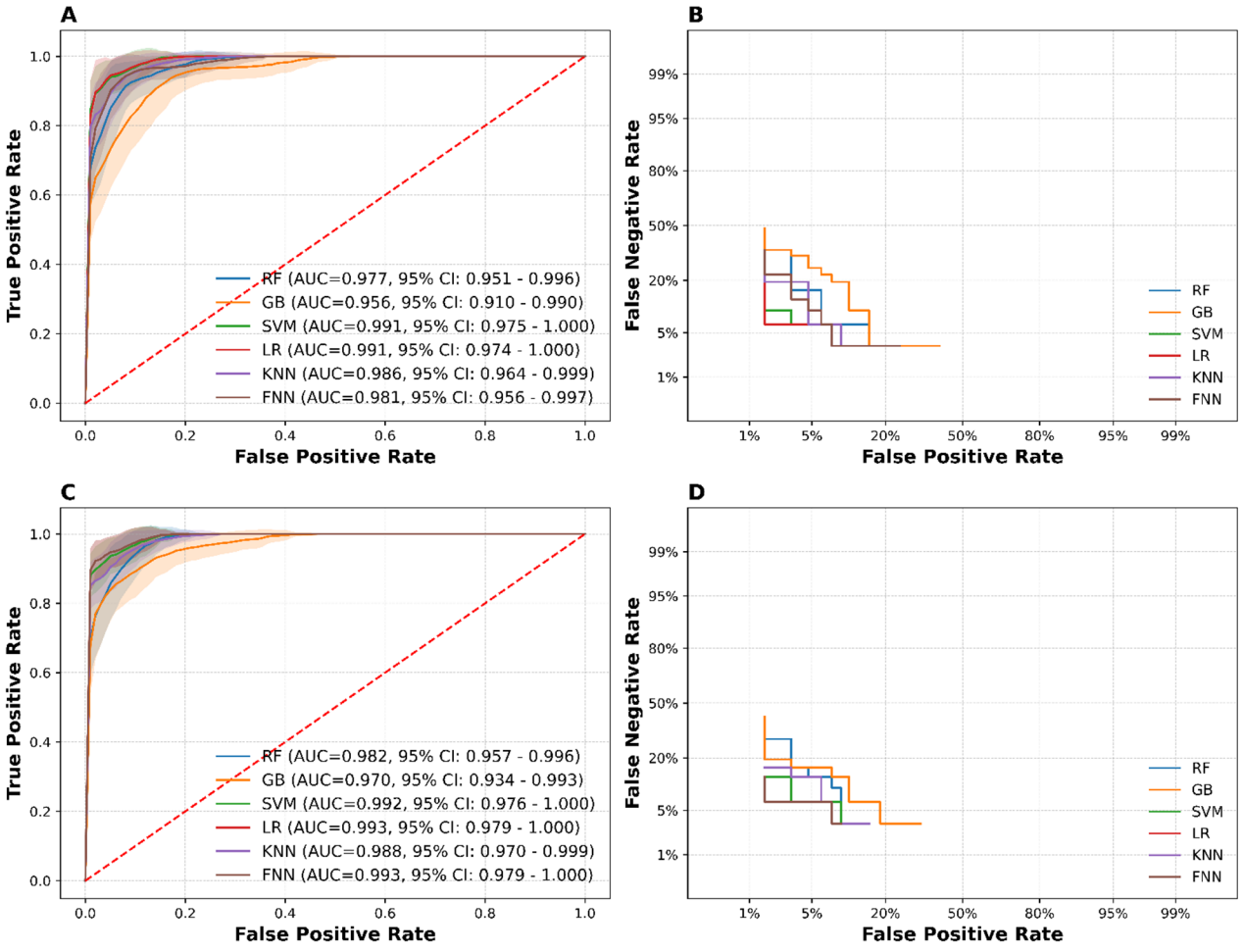

Fig. 4 (**A**) ROC-Curves of PCA-based classification pipelines, (**B**) DET-Curves of PCA-based classification pipelines, (**C**) ROC-Curves of t-SVD-based classification pipelines, (**D**) DET-Curves of the t-SVD-based classification pipelines. LR = Logistic Regression, GB = Gradient Boosting, SVM = Support Vector Machine, RF = Random Forest, KNN = K-Nearest Neighbors, FNN = Feedforward Neural Network

ROC and DET curves for each classification pipeline on the test set. Additionally, the confusion matrices for the PCA- and t-SVD-based classification pipelines, provided in Supplementary Files 2 and 3, allowed for detailed assessment of class-specific errors, including false positives and false negatives. Together, these results offer a comprehensive view of each classification pipeline's behavior, extending beyond aggregate metrics to include both threshold-independent and threshold-sensitive evaluations.

Performance of PCA-based classification pipelines

The PCA-based classification pipelines demonstrated consistently strong performance across nearly all evaluation metrics, confirming the effectiveness of PCA as a dimensionality reduction technique in this clinical context. Among the classifiers, LR emerged as the top performer, achieving the highest balanced accuracy of 0.953 (95% CI: 0.897–0.993) on the test set and 0.859 (95% CI: 0.819–0.899) in stratified 10-fold CV. It also recorded the highest F1 scores (0.958 on the test set, 0.856 in CV), along with exceptional AUC values of 0.991 (test) and 0.965 (CV), demonstrating strong discriminatory power. SVM was another standout model, with AUCs of 0.991 (test) and 0.961 (CV), high sensitivity (0.936 test), and high specificity (0.937 test), confirming its robustness and versatility in classifying clinical outcomes. Although the FNN did not outperform LR on all conventional metrics, it achieved a very high AUC of 0.981 (test) and 0.945 (CV), along with a strong test set sensitivity of 0.934—highlighting its nonlinear decision-making capacity in PCA-transformed feature space.

KNN also demonstrated solid performance, achieving > 0.90 across all test set metrics except for sensitivity, which dropped slightly to 0.872. This decline may be attributed to KNN's reliance on Euclidean distances, which can become less meaningful when neighborhood structures are distorted by linear transformations such

as PCA [64]. Similarly, RF exhibited strong AUCs (0.977 test, 0.965 CV), high specificity (0.969 test), and F1 score (0.926 test), though with slightly reduced sensitivity (0.841 test), reflecting a performance profile similar to KNN. Conversely, GB, while maintaining respectable AUC values (0.956 test, 0.953 CV), demonstrated the lowest sensitivity (0.807 test, 0.797 CV) and balanced accuracy (0.865 test) among the PCA-based pipelines. This suggests limited utility in clinical scenarios where failing to identify true positives—i.e., cases of recurrence—may lead to adverse outcomes. GB also recorded the lowest AUC among all PCA-based pipelines, a result consistent with its suboptimal performance observed in DET curve analyses.

These observations are further supported by Fig. 4A (ROC) and Fig. 4B (DET). The ROC curves for LR, SVM, and FNN are tightly clustered near the top-left corner, signaling excellent classification performance across all thresholds. While ROC analysis captures class separability, Martin et al. (1997) [65] emphasized that DET curves are particularly effective for evaluating classifier performance across varying operational thresholds, offering clearer insights into the trade-off between false positives and false negatives—an essential consideration in decision-critical applications. From a clinical standpoint, DET curves are invaluable: minimizing false negatives is especially critical in recurrence prediction tasks, where undetected cases may lead to delayed intervention and worsened patient outcomes. In this regard, LR and FNN demonstrated the most favorable DET profiles, with curves closest to the origin and stable performance across threshold settings. This indicates they are less sensitive to calibration errors, meaning they can sustain low error rates even when threshold selection is imperfect—a major strength in real-world clinical deployment, where decision thresholds may vary based on risk tolerance, patient stratification, or institutional protocols.

SVM also exhibited favorable DET characteristics but with a slightly more pronounced curve, indicating some susceptibility to threshold shifts. GB and KNN, in contrast, had DET curves that deviated significantly from the origin, reflecting weaker performance in threshold-sensitive scenarios, especially at lower FPRs—conditions that simulate conservative clinical environments where false alarms must be minimized.

In summary, PCA—particularly when paired with LR, SVM, and FNN—yields strong classification pipelines that combine high discriminative power (as seen in ROC analysis) with robust decision threshold behavior (as shown by DET curves), making them well-suited for clinical decision support systems that prioritize safety and diagnostic precision.

Performance of t-SVD-based classification pipelines

The t-SVD-based classification pipelines also demonstrated high performance, with several classifiers performing comparably—and in select cases, even slightly better—than their PCA-based counterparts. LR once again led in balanced accuracy, achieving 0.944 (test) and 0.854 (CV), and it maintained outstanding AUC scores (0.993 test, 0.965 CV) and F1 values (0.958 test, 0.859 CV). These results reinforce LR's capability to perform robustly across different feature representations, making it a reliable choice for clinical implementation. SVM and RF followed closely behind. SVM achieved high AUCs (0.992 test, 0.961 CV) and strong sensitivity (0.928 test), while RF sustained high performance across most metrics, including specificity and precision, suggesting effective control over false positives—a valuable attribute when aiming to reduce overdiagnosis. KNN performed surprisingly well under the t-SVD transformation compared to its PCA-based performance, with AUCs of 0.988 test and 0.952 (CV). However, DET analysis (Fig. 4D) revealed that despite respectable point estimates, its error rates were less consistent across thresholds, again raising concerns over reliability in high-stakes settings.

Interestingly, FNN, while exhibiting a very high AUC (0.993 test), recorded notably low sensitivity (0.644 test, 0.677 CV), suggesting difficulty in capturing positive recurrence cases. However, its perfect specificity (1.000 test) and high precision (0.903) indicate its predictions are highly trustworthy when positive. In clinical settings where overdiagnosis leads to costly or harmful overtreatment, such high-confidence classifiers may still have niche applications. FNN also showed the most favorable DET profile under the t-SVD transformation (Fig. 4D), maintaining low error rates even at conservative thresholds, which may align well with risk-averse clinical scenarios. GB showed improved performance over its PCA variant (balanced accuracy: 0.882 test, 0.882 CV), though it remained among the lower performers in this group. Its AUCs (0.970 test, 0.958 CV) and F1 scores suggest it is a competent but not leading option, with moderate sensitivity and specificity profiles.

The DET curves in Fig. 4D provide important context for clinical application by offering a deeper understanding of each classification pipeline's behavior under varying clinical decision thresholds. For instance, in clinical surveillance of thyroid cancer recurrence, false negatives may delay intervention, while false positives may trigger unnecessary follow-ups or biopsies. In this light, LR and FNN (despite its lower sensitivity) showed the most stable DET profile, making them potentially valuable where low error rates under threshold variability are prioritized. Conversely, SVM and KNN, although strong overall, showed less resilience under extreme threshold conditions, with noticeable FNR spikes at low FPRs. GB again

exhibited an unfavorable DET curve, suggesting limited adaptability in threshold-sensitive environments, which aligns with its known vulnerability to transformed feature spaces.

From a broader perspective, the comparative analysis highlights that PCA-based classification pipelines generally deliver superior and more consistent performance, particularly when paired with LR, SVM, and FNN. PCA's ability to preserve discriminative variance and generate orthogonal, informative features supports a wide range of classifiers—both linear and nonlinear. By contrast, t-SVD offers a distinct advantage in specialized contexts, especially when used with neural networks like FNN, which can exploit latent, low-rank structures and learn complex representations not easily modeled by traditional classifiers.

The inclusion and interpretation of DET curves are particularly impactful in a clinical setting. Unlike ROC curves, which are threshold-independent, DET curves allow researchers and practitioners to assess how classifiers perform when the clinical operating point (i.e., decision threshold) varies. This is crucial in personalized medicine, where different patients or institutions may operate under different tolerances for risk, and it ensures that models are robust to calibration changes. In this context, classifiers with flatter, lower DET curves (such as FNN under t-SVD and LR under PCA) are more clinically reliable, as they minimize the likelihood of performance degradation under uncertainty in threshold selection.

Ultimately, these findings support the clinical viability of dimensionality reduction-based classification pipelines, especially those based on PCA, or t-SVD pairing with LR, SVM and FNN. Future work may explore adaptive or hybrid dimensionality reduction approaches, including ensemble feature selection or manifold learning, to further enhance performance in clinical prediction tasks such as thyroid cancer recurrence monitoring.

Statistical comparison and selection of the best-performing classification pipeline

To assess whether the observed performance differences between PCA- and t-SVD-based classification pipelines were statistically significant ($p < 0.05$), the Wilcoxon signed-rank test was applied to the balanced accuracy scores obtained from stratified 10-fold CV to assess whether the observed performance differences between PCA- and t-SVD-based classification pipelines were statistically significant. As a non-parametric alternative to the paired t-test, the Wilcoxon signed-rank test does not assume normality of the underlying data distribution, making it particularly appropriate for evaluating paired model performance metrics, which are often skewed or non-normally distributed in machine learning experiments. Its application ensures a robust comparison by accounting for both the direction and magnitude of differences between paired observations, thereby enhancing the reliability of performance assessments across classification pipelines [54]. The results (see Table 5) revealed that, for the majority of classifiers—including RF ($p = 0.8316$), GB ($p = 0.0929$), SVM ($p = 0.1730$), and LR ($p = 0.4982$)—the differences in performance between PCA- and t-SVD-based pipelines were not statistically significant. This indicates comparable classifier behavior regardless of the dimensionality reduction technique employed.

Table 5 Statistical comparison of PCA- and t-SVD-based classification pipelines using Wilcoxon signed-rank test

PCA-based pipeline	t-SVD-based pipeline	*p*-value
RF	RF	0.8316
GB	GB	0.0929
SVM	SVM	0.1730
LR	LR	0.4982
KNN	KNN	0.0422*
FNN	FNN	0.0137*

* indicates statistically significant values ($p < 0.05$), LR = Logistic Regression, GB = Gradient Boosting, SVM = Support Vector Machine, RF = Random Forest, KNN = K-Nearest Neighbors, FNN = Feedforward Neural Network

However, statistically significant differences were observed for the KNN and FNN classifiers, with p-values of 0.0422 and 0.0137, respectively. In both cases, the PCA-based pipelines outperformed their t-SVD counterparts in terms of balanced accuracy, suggesting that PCA may provide better feature representation for these specific classifier architectures. These findings underscore the importance of tailoring dimensionality reduction strategies based on the classification algorithm. While PCA and t-SVD generally performed similarly, PCA appears to confer a statistically meaningful advantage for classifiers that rely heavily on geometric distance or gradient-based learning, such as KNN and FNN.

Further supporting these findings, the DET curve analysis revealed that both PCA- and t-SVD-based pipelines generally exhibit low false negative rates. This is especially critical in clinical settings where missing true cases of thyroid cancer recurrence could result in delayed intervention and poorer outcomes. Across both pipelines, LR, RF, SVM, and FNN consistently demonstrated strong classification performance, with GB lagging behind. Interestingly, t-SVD slightly improved GB's performance compared to PCA, indicating that t-SVD may offer greater robustness for certain classifier types with weaker baseline performance.

In the clinical management of thyroid cancer, sensitivity, specificity, and precision are essential metrics. Sensitivity measures the classifier's ability to detect true positives—patients who experience recurrence—and is particularly important to minimize missed diagnoses. For example, LR achieved a near-perfect sensitivity

(> 0.90) on the test set across both pipelines, making it a reliable tool for detecting recurrent cases. KNN also maintained high sensitivity (> 0.87) across both pipelines, demonstrating reliability in early detection. In contrast, FNN showed a substantial sensitivity drop from 0.934 in the PCA-based pipeline to 0.644 in the t-SVD-based pipeline, despite maintaining competitive performance on other metrics. This highlights FNN's sensitivity to the choice of dimensionality reduction technique. However, an exclusive focus on sensitivity may increase false positives, leading to unnecessary interventions. Therefore, the F1 score—which balances sensitivity with precision—and balanced accuracy—which accounts for both sensitivity and specificity—are critical for evaluating clinical model utility. LR demonstrated a high and stable F1 score (0.958) across both pipelines, confirming its balanced ability to identify true positives while minimizing false positives. Similarly, RF, SVM, KNN, and FNN (in the PCA pipeline) all achieved robust F1 scores (> 0.92), with the exception of FNN in the t-SVD pipeline, which dropped to 0.876. These consistent F1 scores reinforce the practical reliability of these models in clinical settings where patient safety and efficient resource use are paramount.

Among all classifiers evaluated, LR emerged as the most suitable classifier for clinical deployment. It offers multiple advantages: (1) probabilistic outputs that allow clinicians to assess recurrence risk on a continuous scale; (2) computational efficiency and ease of implementation for integration into clinical software or web-based platforms; (3) support for incremental learning, enabling timely updates as new patient data become available [42]; and (4) transparent interpretability through its coefficients, which help identify key features influencing recurrence risk [27, 42]. These properties make LR not only effective for prediction but also valuable for clinical decision support and further research into thyroid cancer recurrence.

Taken together, these findings support the use of either PCA- or t-SVD-based classification pipelines in clinical prediction models, with the choice depending on specific performance goals and the characteristics of the classification pipeline. Nonetheless, based on the combination of statistical analysis, performance metrics, and practical considerations, the PCA-based LR classification pipeline was identified as the best performing. It combines high sensitivity, strong F1 score, statistical robustness, and interpretability—qualities that are crucial for real-world healthcare applications. As such, the PCA-based LR classification pipeline was selected for further analysis in this study.

Stratified evaluation of the PCA-based LR classification pipeline across clinically relevant subgroups

To assess the predictive consistency and clinical robustness of the PCA-based LR classification pipeline, we conducted a stratified evaluation across multiple clinically relevant subgroups including age groups, TNM staging, risk levels, and other biologically pertinent factors such as adenopathy, pathology subtype, and tumor focality. The results are summarized in Table 6.

Across age groups, the PCA-based LR classification pipeline demonstrated strong and consistent performance, with balanced accuracy ranging from 0.875 to 0.955 and AUCs exceeding 0.98 in all strata. Notably, the model achieved perfect sensitivity and AUC in the > 60 age group, albeit with reduced specificity due to a smaller test set. In the risk group stratification, the model performed exceptionally in the low-risk group (balanced accuracy = 0.983, F1 = 0.975), but performance declined in the intermediate-risk group (balanced accuracy = 0.600), primarily driven by low specificity (0.200), suggesting overprediction of the positive class. The high-risk group was not evaluable due to the presence of a single outcome class, a limitation stemming from dataset imbalance in that subgroup. Subgroup analysis based on TNM staging revealed variation in model performance depending on stage. For instance, early-stage tumors like T1a and T1b had near-perfect scores, reflecting both high separability and class balance. In contrast, metrics for T3b and T2 showed decreased specificity, hinting at potential ambiguity in these intermediate stages. Notably, T3b, N1b, and certain adenopathy subgroups (e.g., "Left", "Bilateral") showed perfect sensitivity but zero specificity, again due to imbalanced or small sample sizes. These results underscore the importance of larger, more balanced datasets for certain rare or advanced-stage presentations. The model also demonstrated reliable performance in pathological subtypes, particularly papillary and follicular thyroid cancers, where both achieved F1 scores ≥ 0.932 and AUCs of 0.993–1.000. Subtypes such as Hurthle cell and micropapillary carcinoma were underrepresented, limiting generalizability for these rarer histologies. Regarding focality, the classifier maintained high precision and recall across both unifocal and multifocal disease, reflecting stable performance irrespective of tumor multiplicity. Finally, subgroups with no radiotherapy history showed robust metrics (balanced accuracy = 0.937, AUC = 0.991), while the 'Yes' category lacked evaluable cases.

This stratified validation confirms that the PCA-based LR classification pipeline maintains high discriminative ability across most clinically relevant strata, particularly in well-represented categories. It also reveals subgroup-specific limitations tied to class imbalance or

Table 6 Stratified evaluation of PCA-based LR classification pipeline performance across clinically relevant subgroups

Clinically Relevant Subgroup	B. Acc.	F1 score	AUC	Sen.	Spec.	Prec.
Age Group						
45–60	0.955	0.942	0.985	1.000	0.909	0.950
< 45	0.931	0.927	0.991	0.941	0.920	0.933
> 60	0.875	0.913	1.000	1.000	0.750	0.926
Risk Group						
Low	0.983	0.975	0.983	1.000	0.967	0.989
Intermediate	0.600	0.815	0.957	1.000	0.200	0.878
High	N/A	N/A	N/A	N/A	N/A	N/A
TMN Stage						
T2	0.802	0.920	0.969	0.667	0.938	0.928
T3a	0.909	0.924	0.989	1.000	0.818	0.934
T4a	N/A	N/A	N/A	N/A	N/A	N/A
T1a	1.000	1.000	1.000	1.000	1.000	1.000
T1b	0.958	0.934	1.000	1.000	0.917	0.962
T3b	0.500	0.817	0.857	1.000	0.000	0.766
M0	0.934	0.924	0.990	0.960	0.908	0.932
M1	N/A	N/A	N/A	N/A	N/A	N/A
N0	0.920	0.969	0.995	0.857	0.983	0.969
N1b	0.600	0.809	1.000	1.000	0.200	0.875
N1a	0.750	0.733	0.500	1.000	0.500	0.833
Adenopathy Group						
No	0.912	0.957	0.988	0.857	0.967	0.959
Right	0.667	0.800	1.000	1.000	0.333	0.864
Bilateral	0.500	0.817	1.000	1.000	0.000	0.766
Extensive	N/A	N/A	N/A	N/A	N/A	N/A
Left	0.500	0.758	1.000	1.000	0.000	0.694
Pathology Subtype						
Papillary	0.938	0.932	0.993	0.962	0.915	0.936
Follicular	1.000	1.000	1.000	1.000	1.000	1.000
Micropapillary	N/A	N/A	N/A	N/A	N/A	N/A
Hurthel cell	0.667	0.5.	1	1	0.333	0.833
Tumor Focality						
Uni-Focal	0.928	0.966	0.995	0.875	0.980	0.966
Multi-Focal	0.821	0.857	0.984	1.000	0.643	0.889

N/A (Not Available) represent when only one class label is present in the subgroup leading to inability to compute the performance metrics, B. Acc.=Balanced Accuracy, AUC = area under the receiver operating characteristic (ROC) curve, Sen.=Sensitivity, Spec.=Specificity, Prec.=Precision

data sparsity, such as in advanced-stage disease, high-risk tumors, and rarer pathology types.

Explainability and clinical interpretability of the PCA-based classification pipeline using SHAP analysis

To address the critical need for model transparency and clinical interpretability, we conducted a post hoc explainability analysis using SHAP on the best-performing PCA-based LR classification pipeline. The SHAP beeswarm plot (Fig. 5A) and heatmap (Fig. 5B) visualize the contribution of each original clinical feature on the classification pipeline's output, providing both global and local interpretability.

The top-ranking features by impact include Risk group, Response status, T stage, N stage, and overall TNM stage, indicating their substantial influence on the classification pipeline's decision-making. Notably, Risk and Response displayed the largest SHAP value magnitudes, reflecting strong associations with the classification pipeline's predicted probabilities. The SHAP values further reveal directional insights: higher values of Risk and T stage (e.g., more advanced tumor stages) are associated with increased model output toward the predicted class, highlighting their relevance in risk stratification.

Clinically important features like Adenopathy, Thyroid Function, and Pathology subtype also emerged as moderately influential, validating their established roles in thyroid cancer prognosis. On the other hand, demographic and historical variables such as Age, Gender, Focality, and History of Smoking or Radiotherapy showed comparatively lower SHAP value distributions, suggesting that while they are part of the model, their marginal

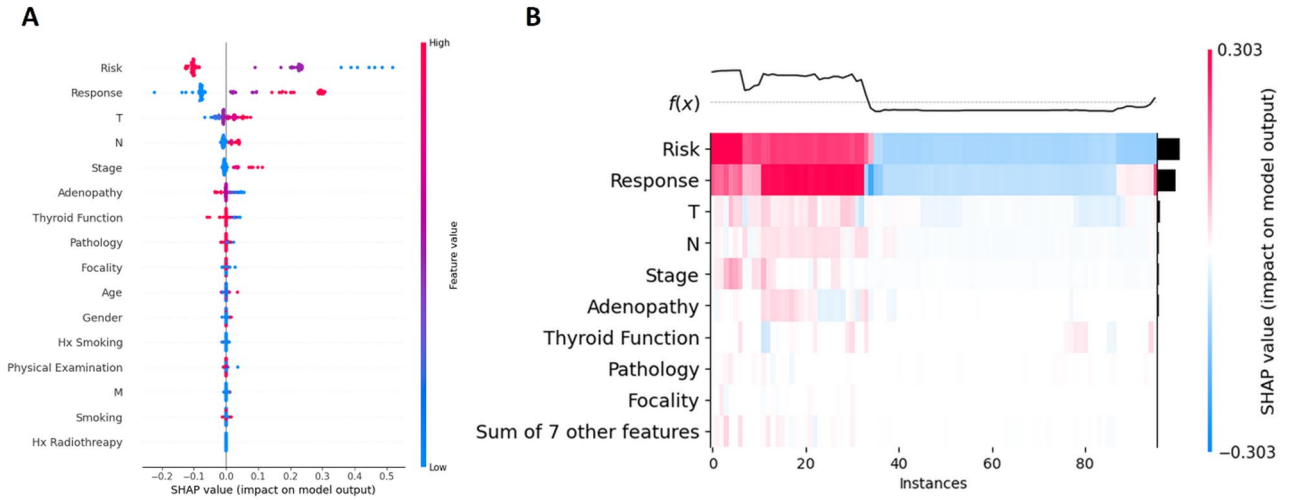

Fig. 5 SHAP-Based Interpretation of Feature Contributions in the Best-Performing PCA-based LR classification pipeline. (**A**) SHAP beeswarm plot illustrating the overall impact, direction, and distribution of each feature's contribution. (**B**) SHAP heatmap displaying the magnitude and direction of individual feature contributions across samples, highlighting patterns relevant to recurrence prediction

contributions to predictions were less pronounced in the studied cohort.

This analysis affirms that PCA-based LR classification pipeline bases its decisions on clinically coherent features and enables clinicians to understand which variables most drive predictions. By mapping model logic back to original clinical factors, we improve transparency, support trustworthiness, and take a step toward ethical, explainable AI deployment in oncology settings.

Comparison with other methods

Compared with existing studies on predicting the likelihood of differentiated thyroid cancer (DTC) recurrence or metastasis in post-treatment patients—including those by Borzooei et al. (2024) [15], Qiao et al. (2024) [66], and Wang et al. (2024) [67]—our study demonstrates several notable strengths and innovations that enhance the accuracy and robustness of predictions:

Advanced dimensionality reduction techniques

Our study employs a comprehensive range of dimensionality reduction techniques to address the curse of dimensionality. We utilized both linear methods, such as PCA and t-SVD, as well as manifold learning techniques like t-SNE and UMAP. This approach allowed us to reduce the dataset to a manageable number of features—three in this case—while retaining those with the highest variance. The application of these methods aligns with recent advancements in data preprocessing for high-dimensional datasets [32].

Rigorous feature engineering and clustering validation

To refine our feature engineering techniques, we implemented k-means clustering on the generated feature sets. We evaluated the clustering results using ARI and V-measure, and assessed the intrinsic properties of each cluster using the silhouette coefficient score. Stratified 10-fold CV was employed to ensure the robustness and reliability of these metrics. This meticulous approach to feature selection and clustering validation reflects best practices in ensuring high-quality feature sets and aligns with methodologies reported by Ester et al. (1996) [68] on clustering evaluation and Hennig (2007) [69] on cluster validation techniques.

Diverse machine learning/deep learning models and comprehensive evaluation metrics

Our study leverages a wide array of classifiers, including a neural network, each with distinct learning behaviors to capture various aspects of the data. We applied diverse classification metrics, including AUC, Balanced Accuracy, Sensitivity, Specificity, Precision, and DET curves. These metrics are particularly valuable for addressing class imbalance and assessing model performance comprehensively. The use of stratified 10-fold CV and testing with an external dataset further ensures that our metrics are generalizable and reflect the true performance of the models on unseen data. This approach is in line with the evaluation practices emphasized by He and Garcia (2009) [48] for dealing with imbalanced datasets and by Sokolova and Lapalme (2009) [70] for evaluating classifier performance.

Robust and reliable performance

Table 7 presents a comparison of our approach with previous studies using selected performance metrics, while Table 8 provides a brief description of the datasets used in the various studies compared with ours. Although

Table 7 Test set performance of PCA- and t-SVD-based classification pipelines compared to previous studies in predicting DTC recurrence

Study	Model	AUC (%)	Sen. (%)	Spec. (%)	Comment
Our Study	SVM	~99.2	87.2–93.6	93.7–98.4	Confirms SVM's effectiveness; aligns with Borzooei et al.
	KNN	~99.0	87.2	95.4	Suggests improved predictive ability with high-variance features.
	RF	~98.0	84.1–87.1	96.9	Consistent with Borzooei et al.; reliable performance.
	FNN	98.1–99.3	64.4–93.4	92.24–1.00	Comparable performance with Borzooei et al.'s ANN model; high spec and sen.
	LR	>99.1	90.4–93.6	96.9–98.4	Superior performance to Wang et al. LR
Borzooei et al. (2024) [15]	SVM	99.71	99.33	97.14	Higher performance than our PCA- and t-SVD-based SVM pipelines (~99.2% AUC).
	KNN	98.44	83	97.14	Our KNN classification pipelines in both pipelines show slightly higher AUC (~99.0%) and sensitivity (87.0%).
	RF	99.38	99.66	94.28	Higher performance than our PCA- and t-SVD-based RF pipelines (>98.0% AUC).
	ANN	99.64	96.6	95.71	High performance comparable to our FNN classification pipeline.
Qiao et al. (2024) [66]	RF	96.0	92.9	N/A	High performance similar to our RF classification pipeline's performance.
Wang et al. (2024) [67]	RF	76.6	0.75.7	68.2	Lower performance than our study; variation may be due to different feature sets.
	LR	73.8	0.86.5	49.5	Lower performance than our LR; variation may be due to different feature sets.
	SVM	75.2	0.56.8	90.3	Lower performance than our SVM but with comparable spec.

Sen.=Sensitivity, Spec.=Specificity, LR=Logistic Regression, GB=Gradient Boosting, SVM=Support Vector Machine, RF=Random Forest, KNN=K-Nearest Neighbors, FNN=Feedforward Neural Network

Table 8 Dataset characteristics across studies compared with the present work

Study	Dataset Description
Borzooei et al. (2024) [15]	Differentiated thyroid cancer dataset from the UCI Machine Learning Repository, consisting of 383 instances and 16 sociodemographic and clinicopathologic features, including age, gender, smoking history, prior radiotherapy, thyroid function, physical exam findings, adenopathy, pathology, focality, risk category, TNM staging, overall stage, and treatment response.
Qiao et al. (2024) [66]	Demographic and clinicopathological data of thyroid cancer patients between 2010 and 2015 extracted from the National Institutes of Health (NIH) Surveillance, Epidemiology, and End Results (SEER) database.
Wang et al. (2024) [67]	Dataset of 2,244 patients who underwent thyroid surgery and radioiodine treatment, including 29 perioperative variables covering demographics, comorbidities, tumor features, lymph node involvement, and metabolic/inflammatory markers.

studies such as Borzooei et al. (2024) [15] demonstrated high sensitivity and specificity using traditional feature sets and models, our study's application of advanced dimensionality reduction and comprehensive feature engineering techniques yielded comparable performance metrics. For example, our PCA-based SVM classification pipeline achieved an AUC of 99.1% and a sensitivity of 93.6%, which is consistent with the high performance of SVM reported by Borzooei et al. [15]. Similarly, the AUC and sensitivity of our PCA- and t-SVD-based RF classification pipelines align with the results reported by Qiao et al. (2024) [66], demonstrating the robustness of Random Forest across various datasets and feature sets.

Clinical implications

The results of this study highlight several important implications for the application of dimensionality reduction and classification techniques in predicting thyroid cancer recurrence. Both PCA and t-SVD proved to be effective dimensionality reduction techniques, facilitating the development of robust classification pipelines. The superior performance of PCA and t-SVD in terms of clustering metrics and PC1 variance suggests their capability to retain significant data structures, which is crucial for accurate classification.

Among the classifiers evaluated, LR consistently delivered the best performance across both PCA- and t-SVD-based pipelines, indicating its robustness and reliability in binary classification tasks related to DTC recurrence prediction. The high precision and balanced accuracy achieved by LR suggest that it is well-suited for clinical applications where accurate prediction and differentiation between recurrence and non-recurrence are critical. RF, SVM, and KNN also demonstrated consistently strong performance across both pipelines. Their robustness may be attributed to their capacity to model complex relationships within the reduced feature space, which is particularly valuable in high-dimensional clinical datasets. Notably, KNN achieved high precision and specificity, indicating its reliability in correctly identifying non-recurrence cases. However, its comparatively lower sensitivity suggests a potential limitation in detecting all true recurrence cases—an important consideration in the clinical management of DTC, where early identification of recurrence is critical for timely intervention. This

underscores the need for further optimization of KNN to enhance its recall without compromising its precision.

FNN also exhibited excellent AUC and precision across both pipelines, especially when paired with t-SVD, reflecting its strong ability to distinguish between recurrence and non-recurrence. However, its markedly reduced sensitivity in the t-SVD configuration compared to PCA highlights a trade-off between discriminative power and the ability to capture all true positive cases. More importantly, this points to FNN's sensitivity to the choice of dimensionality reduction technique—an essential factor when designing clinically deployable models. In the context of DTC recurrence prediction, such variability in sensitivity can impact clinical trust and utility, particularly where false negatives carry significant consequences for patient outcomes. Thus, while FNN may be advantageous in scenarios that prioritize precision and class separability, it requires careful calibration or ensemble integration to ensure clinically acceptable sensitivity.

GB consistently showed the lowest performance across both pipelines, although it exhibited modest improvements with t-SVD. While GB may still offer interpretability and utility in ensemble frameworks, its standalone performance may be insufficient in high-stakes applications such as DTC recurrence prediction, where the cost of missed diagnoses can be substantial.

These findings underscore the importance of selecting appropriate dimensionality reduction and classification techniques to optimize predictive performance. The choice of dimensionality reduction method and model can significantly impact the accuracy and reliability of predictions, with practical implications for improving patient management and treatment strategies in differentiated thyroid cancer.

The classification pipelines developed in this study—particularly the PCA-based LR classification pipeline—hold significant potential for integration into clinical practice, especially for personalized risk assessment. By accurately identifying patients at higher risk of recurrence, this classification pipeline can inform more tailored follow-up protocols and treatment adjustments. For instance, patients flagged by the classification pipelines as high-risk could be scheduled for more frequent monitoring or be considered for adjuvant therapies aimed at reducing recurrence risk.

Limitations and future directions

One notable limitation of this study is its reliance on a single dataset sourced from the UCI Machine Learning Repository. While our classification pipelines—particularly the PCA-based LR classification pipeline—demonstrated strong predictive performance on this dataset, the clinical validity of these findings is inherently constrained by the nature of the data. First, although the dataset offers a well-structured and accessible platform for initial algorithm development and benchmarking, its limited representation in peer-reviewed clinical literature and absence from major indexing platforms such as PubMed raise concerns regarding its widespread acceptance and applicability in clinical research. This warrants the need for caution when extrapolating these findings to real-world clinical settings. However, it is important to note that the dataset was only made publicly available on the UCI Machine Learning Repository on October 30, 2023 [24]. Therefore, the relatively low number of citations or PubMed-indexed studies referencing this dataset is likely attributable to its recent release, rather than a lack of clinical utility or relevance.

Second, although the dataset includes TNM staging information—which can indirectly reflect aspects such as tumor size and regional spread—several critical clinical parameters typically used in thyroid cancer risk stratification, such as extrathyroidal extension, vascular invasion, and completeness of surgical resection [71–73], are either absent or not explicitly documented. While the dataset provides some documentation regarding the timeline of data collection, it lacks details on updates, raising concerns about its alignment with current clinical practices and evolving treatment standards. Future studies should prioritize external validation using independent datasets that incorporate these essential clinical variables and provide temporal context. Such datasets will be instrumental in assessing whether the predictive performance of the PCA-based LR classification pipeline remains robust across different healthcare settings, patient populations, and contemporary clinical guidelines.

Another important direction for future research involves comparing the predictive efficacy of our proposed model against established clinical tools—most notably, the 2015 American Thyroid Association (ATA) guideline risk stratification system, which is widely used in clinical endocrinology to evaluate thyroid cancer recurrence risk [71, 73]. While the dataset used in this study does not explicitly label patients with ATA risk group classifications, it does include multiple features that are critical components of the ATA framework. These include TNM staging (T, N, M), histopathological subtype (Pathology), lymph node involvement (Adenopathy), tumor focality (Focality), and overall cancer stage (Stage). Notably, the dataset al.so contains an overall "Risk" variable, categorized as Low, Intermediate, or High, which approximates ATA-based assessments, derived from tumor size, nodal status, and metastatic spread (see Supplementary File 1 for details). Future work will explore whether more formal mapping of patient records to ATA risk categories can be derived using rule-based clinical logic or consensus expert input. If successful, this would enable direct, head-to-head comparison

between machine learning–based predictions and guideline-based assessments, offering insights into the model's potential for complementing or improving existing clinical decision-making frameworks. Such comparisons would also enhance the model's interpretability and relevance in clinical environments, where guideline adherence remains a key standard of care.

In addition, while this study evaluated several individual classifiers (e.g., LR, RF, SVM, KNN, FNN), future research should explore ensemble learning approaches, including stacking, boosting, or bagging, to further enhance predictive accuracy and model robustness. An intriguing avenue for development would involve building heterogeneous ensemble models where each base learner is trained on features derived from distinct dimensionality reduction techniques (e.g., PCA, t-SVD, UMAP). Such an architecture could exploit the complementary strengths of various feature representations, capturing both linear and nonlinear data structures, and thus potentially improving classification performance across different subgroups of patients.

Lastly, integrating multimodal data—such as genomic, proteomic, or imaging features—with traditional clinical variables may significantly enhance model precision and predictive power. This holistic approach aligns with the goals of precision medicine, where individualized treatment and follow-up strategies are informed by a more comprehensive understanding of each patient's disease biology [74]. As such, expanding the dataset and refining the feature space will be vital to ensure the model's scalability and clinical applicability.

In summary, while our findings present a promising step toward AI-driven clinical decision support in thyroid cancer management, careful validation, model extension, and comparative evaluation against clinical standards are essential next steps for achieving broader clinical adoption.

Conclusions

This study demonstrates that feature engineering techniques, such as PCA and t-SVD, can significantly enhance the performance of classification pipelines in predicting DTC recurrence in post-treatment patients. Classification pipelines incorporating PCA or t-SVD, particularly when paired with models like LR, RF, FNN, SVM, and KNN, showed highly promising results. Among these, LR exhibited the best performance in predicting cancer recurrence and unlike more complex algorithms, such as FNN, LR requires fewer computational resources, ensuring faster predictions and model updates. This approach has the potential to support more effective and personalized treatment strategies, improving patient outcomes by accurately predicting the likelihood of recurrence and enabling timely interventions. In the future, we shall make effort to develop the web-server and standalone software implementing the PCA-based LR classification pipeline utilizing the feature engineering techniques discussed in this study for the prediction of DTC recurrence in post-treatment patients. This tool could bridge the gap between research and clinical practice, ensuring that machine learning advancements are both practical and actionable for healthcare providers. The source codes and generated data, as well as a readme file containing a detailed instructions on how to run the code are freely available from the GitHub link provided here (https://github.com/OnahPmi/Thyroid-Cancer-Recurrence-Prediction-Project).

Abbreviations

DTC	Differentiated thyroid cancer
UCI	University of California Irvine
CV	Cross-validation
ML	Machine learning
PCA	Principal Component Analysis
t-SVD	Truncated Singular Value Decomposition
f-ICA	Fast Independent Component Analysis
NMF	Non-Negative Matrix Factorization
t-SNE	T-distributed Stochastic Neighbor Embedding
Isomap	Isometric Mapping
UMAP	Uniform Manifold Approximation and Projection
LLE	Locally linear embedding
ARI	Adjusted Rand Index
LR	Logistic Regression
GB	Gradient Boosting
SVM	Support Vector Machine
RF	Random Forest
KNN	K-Nearest Neighbors
FNN	Feedforward Neural Network
AUC	Area under the receiver operating characteristic curve
DET	Detection error tradeoff
FPR	False positive rate
FNR	False negative rate
PC1	First principal component
PC2	Second principal component
PC3	Third principal component
ATA	American thyroid association
SHAP	SHapley Additive exPlanations

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03018-3.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements

The authors would like to express their gratitude to Shiva Borzooei and Aidin Tarokhian from Hamadan University of Medical Sciences, Iran, for making the differentiated thyroid cancer dataset publicly available.

Author contributions

E.O. and U.J.E. conceptualized the study. The methodology was developed by E.O., A.S.A., and U.J.E. Validation was carried out by E.O., F.N.K., and U.J.E., while formal analysis was performed by E.O. and U.J.E. The investigation was conducted by E.O. and F.N.K., with resources provided by E.O. Data curation was handled by E.O., U.J.E., and U.G.E. The original draft was prepared by E.O., A.S.A., and K.C.A., with review and editing contributions from K.C.A. and F.N.K.

Visualization was carried out by E.O., U.J.E., and U.G.E. The study was supervised and managed by E.O. and F.N.K. All authors reviewed the manuscript.

Funding
This research received no external funding.

Data availability
Publicly available dataset was analyzed in this study. This data can be found here (https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence). The source codes and generated data are freely available for download from the GitHub link provided here (https://github.com/OnahPmi/Thyroid-Cancer-Recurrence-Prediction-Project).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Nigeria, Nsukka, Enugu State 410001, Nigeria
[2]College of Pharmacy, Ohio State University, Ohio 43210, USA
[3]Department of Pharmacognosy, Faculty of Pharmacy, University of Lagos, Akoka, Yaba, Lagos 101017, Nigeria
[4]School of Pharmacy, University of Pittsburgh, Pittsburgh, PA 15261, USA
[5]Department of Clinical Pharmacy and Pharmacy Management, Faculty of Pharmaceutical Sciences, University of Nigeria, Nsukka, Enugu State 410001, Nigeria
[6]Center for Drug Discovery (UB-CeDD), Faculty of Science, University of Buea, Buea, Cameroon

Received: 23 January 2025 / Accepted: 2 May 2025
Published online: 13 May 2025

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72:7–33. https://doi.org/10.3322/caac.21708.
2. Xi N, Wang L, Yang C. Author correction: improving the diagnosis of thyroid cancer by machine learning and clinical data. Sci Rep. 2022;12:13252. https://doi.org/10.1038/s41598-022-17659-1.
3. Shank JB, Are C, Wenos CD. Thyroid cancer: global burden and trends. Indian J Surg Oncol. 2022;13(1):40–5. https://doi.org/10.1007/s13193-021-01429-y.
4. Książek W. Explainable thyroid cancer diagnosis through two-level machine learning optimization with an improved naked mole-rat algorithm. Cancers (Basel). 2024;16(24):4128. https://doi.org/10.3390/cancers16244128.
5. Christofer JC, Mete O, Baloch ZW. The 2022 WHO classification of thyroid tumors: novel concepts in nomenclature and grading. Endocr Relat Cancer. 2022;30(2):e220293. https://doi.org/10.1530/ERC-22-0293.
6. Miranda-Filho A, Lortet-Tieulent J, Bray F, Cao B, Franceschi S, Vaccarella S, Dal Maso L. Thyroid cancer incidence trends by histology in 25 countries: a population-based study. Lancet Diabetes Endocrinol. 2021;9(4):225–34. https://doi.org/10.1016/S2213-8587(21)00027-9.
7. Jayarangaiah A, Sidhu G, Brown J, Campbell OB, McFarlane SI. Therapeutic options for advanced thyroid cancer. Int J Clin Exp Med. 2019;5:26–34. https://doi.org/10.17352/ijcem.000040.
8. Medas F, Canu GL, Boi F, Lai ML, Erdas E, Calò PG. Predictive factors of recurrence in patients with differentiated thyroid carcinoma: a retrospective analysis on 579 patients. Cancers (Basel). 2019;11(9):1230. https://doi.org/10.3390/cancers11091230.
9. Guo K, Wang Z. Risk factors influencing the recurrence of papillary thyroid carcinoma: a systematic review and meta-analysis. Int J Clin Exp Pathol. 2014;7(9):5393–403.
10. Hakim TJA, Rojas MF, Santivañez JJ, León L, González Devia D. Prognostic factors for recurrence in patients with papillary thyroid carcinoma. Ear Nose Throat J. 2023;1455613231158792. https://doi.org/10.1177/01455613231158792.
11. Alkilany S, Mahfouz E, Mohammed E, Ghazawy E, Abdelgwad Y, Mohamadien N, et al. Recurrence risk in thyroid cancer patients after thyroidectomy. Minia J Med Res. 2024;35:1–10. https://doi.org/10.21608/mjmr.2023.237786.1559.
12. Kim M, Cho SW, Park YJ, Ahn HY, Kim HS, Suh YJ, et al. Clinicopathological characteristics and recurrence-free survival of rare variants of papillary thyroid carcinomas in Korea: a retrospective study. Endocrinol Metab. 2021;36:619–27. https://doi.org/10.3803/EnM.2021.974.
13. Haddad RI, Bischoff L, Ball D, Bernet V, Blomain E, Busaidy NL, Campbell M, Dickson P, Duh QY, Ehya H, Goldner WS, Guo T, Haymart M, Holt S, Hunt JP, Iagaru A, Kandeel F, Lamonica DM, Mandel S, Markovina S, Darlow S. Thyroid carcinoma, version 2.2022, NCCN clinical practice guidelines in oncology. JNCCN. 2022;20:925–51. https://doi.org/10.6004/jnccn.2022.0040.
14. Sarker IH. Machine learning: algorithms, real-world applications, and research directions. SN Comput Sci. 2021;2:1–21. https://doi.org/10.1007/s42979-021-00592-x.
15. Borzooei S, Briganti G, Golparian M, Lechien JR, Tarokhian A. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. Eur Arch Otorhinolaryngol. 2024;281:2095–104. https://doi.org/10.1007/s00405-023-08299-w.
16. Santos MS, Soares JP, Abreu PH, Araújo H, Santos J. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches. IEEE CIM. 2018;13:59–76. https://doi.org/10.1109/MCI.2018.2866730.
17. Clark E, Price S, Lucena T, Haberlein B, Wahbeh A, Seetan R. Predictive analytics for thyroid cancer recurrence: A machine learning approach. Knowledge. 2024;4(4):557–70. https://doi.org/10.3390/knowledge4040029.
18. Elkenawy ESM, Alhussan AA, Khafaga DS, et al. Greylag Goose optimization and multilayer perceptron for enhancing lung cancer classification. Sci Rep. 2024;14:23784. https://doi.org/10.1038/s41598-024-72013-x.
19. Zahraa T, Amel AA, Doaa SK, El-Sayed ME, Ahmed ME. A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its early stages. Biol Signal Process Control. 2025;102:107417. https://doi.org/10.1016/j.bspc.2024.107417.
20. Alkhammash EH, Assiri SA, Nemenqani DM, Althaqafi RMM, Hadjouni M, Saeed F, Elshewey AM. Application of machine learning to predict COVID-19 spread via an optimized BPSO model. Biomimetics. 2023;8(6):457. https://doi.org/10.3390/biomimetics8060457.
21. Elshewey AM, Shams MY, Tawfeek SM, Alharbi AH, Ibrahim A, Abdelhamid AA, et al. Optimizing HCV disease prediction in Egypt: the HyOPTGB framework. Diagnostics. 2023;13(22):3439. https://doi.org/10.3390/diagnostics13223439.
22. Schindele A, Krebold A, Heiß U, Nimptsch K, Pfaehler E, Berr C, Bundschuh RA, Wendler T, Kertels O, Tran-Gia J, Pfob CH, Lapa C. Interpretable machine learning for thyroid cancer recurrence prediction: leveraging XGBoost and SHAP analysis. Eur J Radiol. 2025;186:112049. https://doi.org/10.1016/j.ejrad.2025.112049.
23. Kohavi R. A study of cross-validation and bootstrap for accuracy Estimation and model selection. Proc 14th Int Jt Conf Artif Intell (IJCAI). 1995;2:1137–43.
24. Borzooei S, Tarokhian A. Differentiated Thyroid Cancer Recurrence [dataset]. UCI Machine Learning Repository; 2023. Available from: https://doi.org/10.24432/C5632J
25. Onah E, Uzor PF, Ugwoke IC, Eze JU, Ugwuanyi ST, Chukwudi IR, Ibezim A. Prediction of HIV-1 protease cleavage site from octapeptide sequence information using selected classifiers and hybrid descriptors. BMC Bioinformatics. 2022;23:466. https://doi.org/10.1186/s12859-022-05017-x.
26. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat. 2010;2(4):433–59.
27. Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer Series in Statistics; 2002. pp. 338–9.
28. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw. 2000;13(4–5):411–30.
29. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91. https://doi.org/10.1038/44565.
30. van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
31. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000;290(5500):2319–23. https://doi.org/10.1126/science.290.5500.2319.

32. McInnes L, Healy J, Melville JUMAP. Uniform Manifold Approximation and Projection for Dimension Reduction. J Open Source Softw. 2018;3(29):861. Available from: https://arxiv.org/abs/1802.03426
33. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000;290(5500):2323–6. https://doi.org/10.1126/science.290.5500.2323.
34. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218. https://doi.org/10.1007/BF01908075.
35. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846–50. https://doi.org/10.1080/01621459.1971.10482356.
36. Rosenberg A, Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL); 2007. pp. 410–20. Available from: https://aclanthology.org/D07-1043.
37. Rousseeuw PJ, Silhouettes. A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
38. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken, NJ: Wiley; 2013.
39. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
40. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. pp. 785–94.
41. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
42. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
43. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
44. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
45. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13:281–305.
46. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognit. 2015;48:2839–46.
47. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: Proceedings of the 20th International Conference on Pattern Recognition; 2010 Aug 23–26; Istanbul, Turkey. IEEE; 2010. pp. 3121–4.
48. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.
49. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74.
50. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7):1145–59.
51. Forman G. An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res. 2003;3:1289–305.
52. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2(1):37–63.
53. Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton: Chapman & Hall/CRC; 1993. pp. 45–57.
54. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1945;1(6):80–3. https://doi.org/10.2307/3001968.
55. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
56. Python Software Foundation. Python Language Reference, Version 3.8.10 [Internet]. 2023 [cited YYYY MMM DD]. Available from: https://www.python.org/
57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30. Available from: https://scikit-learn.org/stable/about.html
58. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
59. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with numpy. Nature. 2020;585:357–62. https://doi.org/10.1038/s41586-020-2649-2.
60. McKinney W. Data Structures for Statistical Computing in Python. In: Proceedings of the 9th Python in Science Conference; 2010 Jun 28–Jul 3; Austin, TX. pp. 56–61.
61. Hunter JD, Matplotlib. A 2D graphics environment. Comput Sci Eng. 2007;9(3):90–5. https://doi.org/10.1109/MCSE.2007.55.
62. Waskom ML, Seaborn. Statistical data visualization. J Open Source Softw. 2021;6(60):3021. https://doi.org/10.21105/joss.03021.
63. Berrar D. Cross-validation. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of bioinformatics and computational biology. Volume 1. Oxford: Academic; 2019. pp. 542–5.
64. Aggarwal CC, Hinneburg A, Keim DA. On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche J, Vianu V, editors. Database Theory — ICDT 2001. Lecture Notes in Computer Science. Volume 1973. Berlin, Heidelberg: Springer; 2001. pp. 420–34. https://doi.org/10.1007/3-540-44503-X_27.
65. Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M. The DET curve in assessment of detection task performance. In: Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997); 1997; Rhodes, Greece. pp. 1895–8. https://doi.org/10.21437/Eurospeech.1997-504
66. Qiao L, Li H, Wang Z, Sun H, Feng G, Yin D. Machine learning based on SEER database to predict distant metastasis of thyroid cancer. Endocrine. 2024;84:1040–50. https://doi.org/10.1007/s12020-023-03657-4.
67. Wang H, Zhang C, Li Q, Tian T, Huang R, Qiu J, et al. Development and validation of prediction models for papillary thyroid cancer structural recurrence using machine learning approaches. BMC Cancer. 2024;24:427. https://doi.org/10.1186/s12885-024-12146-4.
68. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96); 1996. pp. 226–31.
69. Hennig C. Cluster-wise assessment of cluster stability. Comput Stat Data Anal. 2007;52:258–71.
70. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45:427–37. https://doi.org/10.1016/j.ipm.2009.03.002.
71. Xu B, Ghossein RA. Crucial parameters in thyroid carcinoma reporting - challenges, controversies and clinical implications. Histopathology. 2018;72(1):32–9. https://doi.org/10.1111/his.13335.
72. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American thyroid association guidelines task force on thyroid nodules and differentiated thyroid Cancer. Thyroid. 2016;26(1):1–133. https://doi.org/10.1089/thy.2015.0020.
73. Tuttle RM, Alzahrani AS. Risk stratification in differentiated thyroid cancer: from detection to final Follow-Up. J Clin Endocrinol Metab. 2019;104(9):4087–100. https://doi.org/10.1210/jc.2019-00177.
74. Jameson JL, Longo DL. Precision medicine–personalized, problematic, and promising. N Engl J Med. 2015;372(23):2229–34. https://doi.org/10.1056/NEJMsb1503104.

Publisher's note