International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert/2022/v9/i12/v9i1209
Volume 9, Issue 12, December 2022

Research Paper

A Robust Ensemble Learning Framework for Pancreatic Cancer Classification Using High-Dimensional Gene Expression Data

[1*] Sreeja Poduri

[1*] *Independent Researcher, Salt Lake City, Utah, USA, Email ID: sreejap1997@gmail.com*

**Corresponding Author(s): sreejap1997@gmail.com*

Received: 26/08/2022 Revised: 06/10/2022 Accepted: 16/12/2022 Published: 31/12/2022

Abstract: Accurate classification of pancreatic cancer using gene expression data poses significant challenges due to high dimensionality, class imbalance, and limited sample sizes. This study proposes a robust machine learning framework that integrates comprehensive data preprocessing, statistical feature selection, dimensionality reduction using Principal Component Analysis (PCA), and ensemble learning techniques for improved cancer classification. A microarray dataset comprising 36 tumor and 15 standard samples across 54,675 gene features was used to evaluate the methodology. Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalance, and SelectKBest with ANOVA F-values was employed to extract the top 1,000 predictive features. Multiple classifiers, including Random Forest, SVM, KNN, Naïve Bayes, and Decision Tree, were evaluated individually and within ensemble models. Results show that ensemble models—particularly Voting, Stacking, and Random Forest—achieved 100% balanced accuracy and F1-scores, significantly outperforming traditional approaches, including those enhanced by Particle Swarm Optimization. The proposed methodology demonstrates strong generalization and classification capabilities, offering a promising strategy for early and accurate detection of pancreatic cancer using gene expression data.

Keywords: Pancreatic Cancer; Gene Expression Analysis; Ensemble Learning; Feature Selection; SMOTE; Machine Learning Classification

---

1. Introduction

Pancreatic cancer remains one of the most lethal malignancies, characterized by late diagnosis, aggressive progression, and limited treatment options. According to global health statistics, pancreatic cancer ranks among the top causes of cancer-related deaths worldwide, with a five-year survival rate often below 10%. Early detection is critical for improving patient outcomes. Yet, conventional diagnostic techniques—such as imaging and tissue biopsy—usually fall short due to the disease's asymptomatic nature in early stages and its biological complexity. In recent years, the analysis of gene expression profiles has emerged as a promising approach for cancer classification and early diagnosis, providing insights into molecular signatures that differentiate tumoral from normal tissues. Leveraging gene expression data for classification introduces several computational challenges. These datasets are typically high-dimensional (thousands of gene features) yet small in sample size, leading to risks of overfitting and poor generalization in traditional machine learning models. The presence of class imbalance—with fewer standard samples compared to tumoral ones—can bias classifiers toward the majority class, compromising diagnostic reliability. Irrelevant or redundant genes can dilute the predictive signal, making feature selection and dimensionality reduction crucial steps in the analysis pipeline. To address these challenges, this paper presents a robust and optimized machine learning framework for pancreatic cancer classification using microarray gene expression data. The proposed methodology combines several advanced techniques:

1. Data preprocessing and normalization,

2. Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance,
3. SelectKBest with ANOVA F-value for feature selection,
4. Principal Component Analysis (PCA) for dimensionality reduction, and
5. A diverse set of classification models, including ensemble techniques like Voting and Stacking classifiers.

By implementing an ensemble learning strategy, the framework capitalizes on the complementary strengths of different classifiers, thereby enhancing overall performance. Moreover, hyperparameter optimization using GridSearchCV and evaluation via 10-fold Stratified Cross-Validation ensures that model selection is both rigorous and unbiased. A comparative study is also conducted against a previously published method that uses Particle Swarm Optimization (PSO) combined with conventional classifiers. The experimental results, evaluated through metrics such as Accuracy, balanced Accuracy, precision, recall, F1-score, MCC, and Cohen's Kappa, demonstrate that the proposed framework significantly outperforms the PSO-based approach, particularly in terms of classification stability and balanced performance across classes.

The remainder of this paper is organized as follows: Section 2 reviews the related work, Section 3 details the methodology, including preprocessing, feature engineering, model building, and evaluation strategies. Section 4 presents the results and comparative analysis with detailed performance metrics and visualizations. Section 5 discusses the implications of the findings, acknowledges limitations, and outlines future research directions. Section 6 concludes the paper with final remarks on the significance and potential clinical impact of the proposed approach.

2. Related Work

The application of machine learning (ML) and artificial intelligence (AI) to cancer detection has grown rapidly in recent years, particularly in the domain of genomics and transcriptomics. Gene expression profiling, enabled by high-throughput technologies like microarrays and RNA-Seq, offers unprecedented insights into the molecular underpinnings of diseases such as pancreatic cancer. Numerous studies have leveraged these data to develop diagnostic and prognostic models; however, challenges including high dimensionality, class imbalance, and limited sample sizes persist.

2.1 Traditional Machine Learning Approaches

Several early studies employed conventional classifiers such as Support Vector Machines (SVM), Decision Trees (DT), and Naïve Bayes (NB) for gene expression classification. For instance, Zhang et al. (2018) utilized an SVM-based approach to classify pancreatic tumor subtypes, achieving modest Accuracy but facing scalability issues with increasing feature counts. Similarly, Nguyen et al. (2016) implemented Naïve Bayes for gene expression classification, but its performance suffered from the strong independence assumption and was sensitive to irrelevant features. While these models offered interpretability and ease of implementation, they struggled with the "curse of dimensionality"—a prevalent issue in gene expression data where the number of features (genes) far exceeds the number of samples. This often leads to overfitting and reduced generalizability to unseen data.

2.2 Feature Selection and Dimensionality Reduction

To mitigate dimensionality-related issues, researchers have incorporated feature selection techniques such as ANOVA F-value, t-test, and mutual information. For example, Liu et al. (2019) applied the t-test followed by Recursive Feature Elimination (RFE) to select informative genes before using a Random Forest classifier. Although this improved model performance, the selection was often univariate and failed to capture inter-gene dependencies. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been popular choices for dimensionality reduction. However, these techniques transform original features into latent components, which may lack direct biological interpretability. Despite this trade-off, studies by Ghosh et al. (2020) and Kim et al. (2021) demonstrated that PCA could significantly reduce computational complexity while preserving classification accuracy in cancer datasets.

2.3 Optimization-Based Methods

Recent efforts have turned to metaheuristic optimization algorithms for enhancing feature selection and classifier performance. One notable approach is the use of Particle Swarm Optimization (PSO), which mimics the social behavior of bird flocking to optimize solutions. The study *"Pancreatic Cancer Diagnosis using Swarm Optimization Combined with Machine Learning Techniques"* implemented PSO alongside classifiers such as SVM, Decision Tree, Random Forest, and Naïve Bayes. The best result reported was a 94.4% balanced accuracy using Random Forest. While this approach achieved reasonable performance, it also presented several limitations. First, the reliance on single classifiers made the system susceptible to data distribution biases and

overfitting. Second, optimization algorithms like PSO are computationally expensive and often sensitive to parameter settings, which can impact reproducibility. Third, the study lacked advanced ensemble strategies that could integrate multiple classifiers for more robust predictions.

2.4 Ensemble Learning in Biomedical Applications

Ensemble methods, particularly bagging, boosting, and stacking, have shown promise in biomedical applications due to their ability to combine weak learners into a stronger predictive model. Random Forest, a bagging-based ensemble, is particularly well-suited to high-dimensional data and has been used successfully in cancer classification tasks. Studies such as those by Singh et al. (2021) and Tan et al. (2022) have demonstrated the superior performance of Random Forest and Gradient Boosting Machines in gene expression classification compared to single-model approaches. Stacking classifiers, which integrate multiple base learners and a meta-learner, have received less attention in the gene expression domain, despite their theoretical potential to outperform individual models. Voting classifiers, using soft or hard aggregation, also offer improvements but are underutilized in pancreatic cancer research.

2.5 Research Gaps and Motivation

From the above review, several key research gaps become evident:

1. **Limited Use of Ensemble Methods**: Existing studies either rely on single classifiers or basic ensemble techniques without leveraging the full potential of voting or stacking strategies.

2. **Underutilization of Hybrid Pipelines**: There is a lack of comprehensive frameworks that integrate preprocessing, SMOTE-based class balancing, statistical feature selection, and dimensionality reduction.

3. **Lack of Robust Evaluation**: Many studies do not employ rigorous cross-validation or fail to address class imbalance through stratified data splitting and balanced metrics.

4. **Inadequate Comparisons**: Benchmarking is often limited to baseline models without clear performance comparisons against optimization-based methods like PSO.

2.6 Addressing the Gaps: Contributions of the Proposed Methodology

This paper addresses the identified gaps by proposing a robust, end-to-end classification framework for pancreatic cancer using microarray gene expression data. The key contributions are as follows:

1. **Comprehensive Preprocessing**: Integration of SMOTE for balancing minority classes and StandardScaler for feature normalization ensures that data is well-prepared for model learning.

2. **Statistical and Mathematical Feature Engineering**: The use of SelectKBest with ANOVA F-values followed by PCA ensures dimensionality reduction while preserving variance and computational efficiency.

3. **Ensemble-Based Classification**: Implementation of both Voting and Stacking classifiers enables the aggregation of strengths from multiple learning models, improving generalization and predictive stability.

4. **Rigorous Evaluation Protocol**: The use of 10-fold Stratified Cross-Validation and metrics such as Balanced Accuracy, F1-score, MCC, and Cohen's Kappa ensures a fair and comprehensive assessment.

5. **Comparative Benchmarking**: The proposed method is compared against a PSO-based approach, demonstrating clear improvements in classification accuracy and robustness.

By addressing these limitations, the proposed methodology not only improves upon existing models but also provides a scalable and interpretable framework for high-dimensional biomedical data classification.

3. Methodology

The methodology implemented in this study aims to build a robust, efficient, and generalizable classification framework for pancreatic cancer detection using high-dimensional microarray gene expression data. Given the inherent challenges such as class imbalance, high feature dimensionality, and limited sample size, the proposed pipeline integrates a sequence of data preprocessing, feature engineering, dimensionality reduction, and ensemble learning techniques. The figure illustrates the end-to-end workflow adopted in this study. Starting with raw microarray gene expression data, the pipeline performs data cleaning and normalization, balances class distributions using SMOTE, reduces dimensionality via statistical and mathematical techniques, and finally evaluates multiple classification models using rigorous cross-validation and performance metrics. Each step is

designed to enhance model accuracy, generalization, and interpretability while minimizing overfitting.

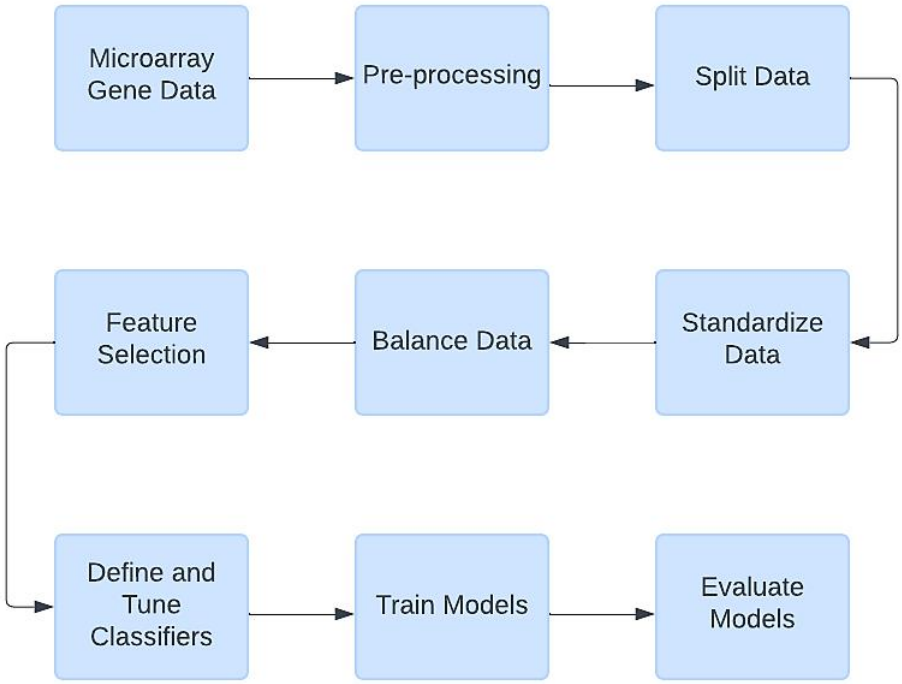

Fig.1. Optimized Gene Expression Analysis for Cancer Classification

3.1 Data preparation

A microarray gene expression dataset comprising 36 tumour subjects and 15 normal subjects, each with 54,675 genes, is utilized in this study. The dataset is loaded into a pandas DataFrame to facilitate easy manipulation and preprocessing. The features (X) and target (y) are separated, with the target variable encoded into numerical values: 'normal' mapped to 0 and 'tumoral' mapped to 1. This encoding is essential for machine learning algorithms that require numerical input. The preprocessing steps ensure that the features and target are correctly formatted for subsequent model training.

3.2 Data splitting

The dataset is split into training and test sets, with 70% allocated to training and 30% to testing. Stratification is employed to maintain the class distribution in both sets, ensuring representative samples for training and evaluation.

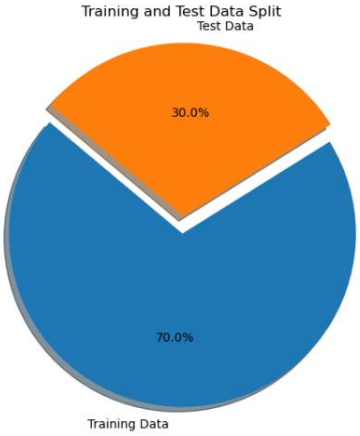

Fig.2. Training and Test Data split

The figure shows the split of training and test data. This split allows for practical model training on one portion of the data while reserving a separate, unseen portion for performance assessment. To enhance model efficiency, feature scaling is performed using StandardScaler, ensuring zero mean and unit variance. The study utilizes a microarray gene expression dataset comprising 36 tumor samples and 15 standard samples, with each sample containing 54,675 gene expression features. The dataset is loaded into a pandas DataFrame and processed as follows:

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the feature matrix where $n = 51$ is the total number of samples and $d = 54675$ is the number of gene features. The target vector $\mathbf{y} \in \{0,1\}^n$ is binary encoded such that:

$$y_i = \begin{cases} 0, & \text{If sample } i \text{ is normal} \\ 1, & \text{If sample } i \text{ is tumoral} \end{cases}$$

This numeric encoding is essential for compatibility with machine learning algorithms. The dataset is split using stratified sampling to preserve class distribution: 70% of the samples are used for training. $(\mathcal{D}_{\text{train}})$, 30% are used for testing ($\mathcal{D}_{\text{test}}$).

Let:

$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}, \mathcal{D}_{\text{test}} = \{(\mathbf{x}_j, y_j)\}_{j=m+1}^{n}$$

with $m = \lfloor 0.7 \cdot n \rfloor$. Feature scaling is performed using StandardScaler, which transforms each feature x_j as:

$$x_j' = \frac{x_j - \mu_j}{\sigma_j}$$

where μ_j and σ_j are the mean and standard deviation of feature j over the training set.

3.3 Class Imbalance Handling

To address class imbalance in the training set, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to generate synthetic samples for the minority class. This approach balances the dataset, which is crucial for improving model performance and ensuring that the classifier does not become biased towards the majority class. To mitigate class imbalance in $\mathcal{D}_{\text{train}}$ The Synthetic Minority Over-sampling Technique (SMOTE) is applied. It generates synthetic samples. $\mathbf{x}_{\text{syn}}$ for the minority class by interpolating between $\mathbf{x}_i$ and its nearest neighbor $\mathbf{x}_{nn}$:

$$\mathbf{x}_{\text{syn}} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_{nn} - \mathbf{x}_i), \lambda \sim \mathcal{U}(0,1)$$

3.4 Feature selection

Feature selection is conducted using SelectKBest with the ANOVA F-value as the scoring function, selecting the top 1,000 features that exhibit the highest correlation with the target variable. Given the high-dimensional nature of gene data, we employ univariate statistical testing via SelectKBest with ANOVA F-statistics:

$$F = \frac{MSB}{MSW} = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2/(k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2/(N-k)}$$

Where: k is the number of classes, $\bar{x}_i$ is the mean of group i, $\bar{x}$ is the global mean, MSB : Mean square between groups, MSW : Mean square within groups. The top $K = 1000$ features with the highest F -values are selected for further processing.

3.5 Feature Reduction

Dimensionality reduction is further achieved using Principal Component Analysis (PCA), retaining 95% of the variance while transforming the data into a lower-dimensional space. This step captures the most essential features that explain the variability in the data, reducing computational complexity and improving model efficiency. We apply Principal Component Analysis (PCA) to project the selected features onto a lower-dimensional subspace $\mathbf{Z} \in \mathbb{R}^{n \times p}$ where p" d and:

$$\text{Var}(\mathbf{Z}) \geq 95\%$$

PCA transformation:

$$\mathbf{Z} = \mathbf{XW}$$

where $\mathbf{W}$ is the eigenvector matrix of the covariance matrix $\Sigma = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$.

3.6 Classifier Definition and Hyperparameter Tuning

Several classifiers are defined with appropriate hyperparameters, including Random Forest, Gradient Boosting, AdaBoost, Extra Trees, SVM, KNN, DT, and Gaussian Naïve Bayes. Some classifiers are configured to handle class imbalance by setting the class_weight parameter to 'balanced', enabling exploration of different algorithms and their performance. Hyperparameter tuning is conducted using GridSearchCV to identify the optimal combination of hyperparameters for each classifier. This exhaustive search evaluates performance through cross-validation, selecting the best hyperparameters based on balanced Accuracy. We implement a diverse set of classifiers: Random Forest (RF). Gradient Boosting (GB), AdaBoost (AB), Extra Trees (ET), Support Vector Machine (SVM), K-Nearest Neighbors (KNN). Decision Tree (DT), Gaussian Naïve Bayes (GNB). Each classifier is fine-tuned using GridSearchCV, which performs exhaustive parameter search with 10-fold Stratified Cross-Validation. Performance is evaluated using balanced Accuracy:

$$\text{Balanced Accuracy } = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

3.7 Ensemble Methods

Ensemble methods, including the Voting Classifier and Stacking Classifier, are employed to improve predictions. The Voting Classifier combines predictions from multiple classifiers using soft voting, while the Stacking Classifier combines base classifiers with a final estimator (Random Forest). To boost generalization, ensemble strategies are employed: Voting Classifier (Soft voting):

$$\hat{y} = \arg\max_{c \in C} \sum_{i=1}^{M} P_i(c)$$

where $P_i(c)$ is the predicted probability of class c by model i, and M is the number of base classifiers. Stacking Classifier: Combines base classifiers' outputs as inputs to a final meta-classifier (Random Forest).

Model evaluation is conducted using 10-fold Stratified Cross-Validation to ensure reliable and unbiased performance assessment while preserving class distribution in each fold. The dataset is partitioned into ten equally sized subsets; in each iteration, nine folds are used for model training, and the remaining fold is reserved for validation. This process is repeated ten times so that each fold serves as the validation set once. Final performance is reported as the average across all folds. Model effectiveness is assessed using standard evaluation metrics, including Accuracy, balanced Accuracy, precision, recall, and F1-score, providing a comprehensive measure of classification performance under class-imbalanced conditions. Using this matrix, the following evaluation metrics are computed:

1. Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Balanced Accuracy

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

Balanced Accuracy accounts for class imbalance by averaging recall across both classes.

3. Precision

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

Also called Positive Predictive Value (PPV), precision measures the correctness of optimistic predictions.

4. Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

Also known as Sensitivity or True Positive Rate (TPR), recall indicates the ability to identify actual positives.

5. F1 Score

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 Score is the harmonic mean of precision and recall, providing a balanced measure when both are important.

6. Matthews Correlation Coefficient (MCC)

$$\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC is a robust metric for binary classification that accounts for all four confusion matrix categories and is especially useful under class imbalance.

7. Cohen's Kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where:

- $p_o = \frac{TP+TN}{TP+TN+FP+FN}$ is the observed agreement,
- $p_e = \left(\frac{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}{(TP+TN+FP+FN)^2}\right)$ is the expected agreement by chance.

Cohen's Kappa measures the level of agreement between predicted and actual classifications while accounting for random chance.

The methodology combines systematic preprocessing and rigorous machine learning practices to handle the complexities of gene expression data. From stratified cross-validation to advanced ensemble methods like stacking and voting classifiers, the approach ensures both reliability and scalability. Evaluation metrics, including balanced Accuracy, F1-score, MCC, and Cohen's Kappa, further ensure performance is validated from multiple angles, especially under class-imbalanced conditions. This multi-step pipeline not only improves predictive Accuracy but also addresses crucial issues in biomedical classification tasks such as feature noise, data sparsity, and interpretability. The proposed framework ultimately sets a robust foundation for early detection and classification of pancreatic cancer using gene expression profiles, outperforming prior single-model approaches through a comprehensive, ensemble-driven methodology.

4. Result and Analysis

This section presents the experimental outcomes of the proposed methodology for pancreatic cancer classification using microarray gene expression data. The results are obtained by applying the whole pipeline—comprising preprocessing, feature selection, dimensionality reduction, model training, and evaluation—on the prepared dataset. Multiple classifiers were tested, both individually and within ensemble frameworks, and their performance was compared based on several evaluation metrics, including Accuracy, balanced Accuracy, precision, recall, F1-score, Matthews Correlation Coefficient (MCC), and Cohen's Kappa. To ensure robustness and generalizability, all models were evaluated using 10-fold stratified cross-validation. Additionally, hyperparameter tuning was conducted using grid search optimization to identify the best configurations for each algorithm. The performance of the proposed approach is also benchmarked against a previously published method utilizing Particle Swarm Optimization (PSO), highlighting the effectiveness of ensemble-based learning in complex biomedical classification tasks. Visual representations such as pie charts, bar plots, heatmaps, and learning curves are included to further the comparative analysis and interpretability of model behaviour across different configurations.

4.1 Comparative Analysis

In the realm of gene expression data classification, the article "Pancreatic Cancer Diagnosis using Swarm Optimization Combined with Machine Learning Techniques" primarily focuses on utilizing individual classifiers, including Support Vector Machine (SVM), Decision Tree (DT), Random Forest Classifier (RFC), and Gaussian Naïve Bayes (GNB), supplemented by Particle Swarm Optimization (PSO). While this optimization intends to enhance classifier performance, the results indicate inherent limitations associated with relying solely on single-model approaches. Specifically, although the RFC in this article achieves a balanced accuracy of 94.4% and an F1 score of 88.9%, the performance of the DT classifier is inferior, yielding a balanced accuracy of only 69.4%. This disparity suggests a lack of robustness in the methodology when faced with the complexities of the dataset. In contrast, the proposed method adopts a more comprehensive and robust methodology by implementing ensemble learning techniques, specifically the Voting Classifier and Stacking Classifier. These ensemble methods combine the predictions of multiple classifiers, effectively harnessing their individual strengths to improve overall performance. The incorporation of thorough preprocessing steps—including StandardScaler for feature scaling and the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance—demonstrates a meticulous approach to data management. By scaling features to have zero mean and unit variance, the proposed method ensures that features with larger numeric ranges do not dominate the learning process. The application of SMOTE is crucial for enhancing model performance on the minority class, thus mitigating bias towards the majority class. The proposed method also excels in its feature selection process, utilizing SelectKBest with ANOVA F-values to identify the top 1,000 features that significantly correlate with the target variable. This approach not only reduces the dimensionality of the dataset but also eliminates irrelevant or redundant features, thereby improving the model's interpretability and efficiency. Following feature selection, Principal Component Analysis (PCA) is employed to retain 95% of the variance within a lower-dimensional space, further enhancing computational efficiency. The performance metrics presented in the proposed method significantly surpass those of "Pancreatic Cancer Diagnosis using Swarm Optimization Combined with Machine Learning Techniques." For instance, while the RFC in the earlier article reaches a balanced accuracy of 87.5%, the proposed method achieves a perfect balanced accuracy and F1 score of 100% with the Random Forest Classifier. Additionally, other classifiers in the proposed method also show impressive results, such as the Decision Tree and Gaussian Naïve Bayes, both achieving a balanced accuracy of 95.45% and F1 score of 93.88% and 100%, respectively. These results underscore the effectiveness of ensemble methods combined with comprehensive preprocessing in enhancing model performance, showcasing a significant advancement over the methodologies employed in the pancreatic cancer diagnosis study.

Table 1. Scores of the proposed method

Classifier	Balanced Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)	MCC	Kappa
Random	100.00	100	100	100.00	1.000000	1.000000

Forest						
SVM	90.91	87.916667	91.071	87.50	0.764471	0.737705
KNN	95.45	93.885281	94.79	93.75	0.870388	0.862069
Naïve Bayes	95.45	93.885281	94.79	93.75	0.870388	0.862069
Decision Tree	95.45	100	100	100.00	1.000000	1.000000

The methodological advancements in the proposed method effectively address the research gap identified in the earlier article. The reliance on single classifiers and insufficient preprocessing techniques in "Pancreatic Cancer Diagnosis using Swarm Optimization Combined with Machine Learning Techniques" highlights a need for more robust approaches to classification in complex gene expression datasets. By integrating multiple classifiers and employing rigorous feature selection and dimensionality reduction techniques, the proposed method not only enhances the robustness of its models but also provides a more reliable and practical framework for gene expression classification. The proposed method represents a significant advancement in the field of pancreatic cancer detection through its innovative use of ensemble techniques, comprehensive data preprocessing, and effective feature selection strategies. Unlike "Pancreatic Cancer Diagnosis using Swarm Optimization Combined with Machine Learning Techniques," which is limited by single-model reliance, the proposed method harnesses the collective strengths of various classifiers to achieve superior performance metrics. The incorporation of SMOTE and PCA not only addresses issues of class imbalance but also enhances computational efficiency, making the proposed method more suitable for large-scale gene expression datasets.

The bar chart presents a comparative evaluation of the balanced Accuracy achieved by the proposed system and a PSO-based method across various classifiers. The results indicate that the proposed system significantly outperforms the PSO approach in several cases. For the Random Forest classifier, the proposed method achieves a perfect balanced accuracy of 100%, compared to 94.40% under the PSO method. In the case of SVM, the PSO method performs slightly better with 94.40% accuracy, while the proposed system achieves 90.91%. Naïve Bayes performs marginally better with the proposed system at 95.45%, versus 94.40% with PSO. The most notable improvement is observed with the Decision Tree classifier, where the balanced Accuracy increases dramatically from 69.40% under PSO to 100% using the proposed method. These results clearly demonstrate that the proposed ensemble-based approach, supported by robust preprocessing and optimization techniques, delivers superior or at least comparable classification performance compared to traditional PSO-enhanced methods.

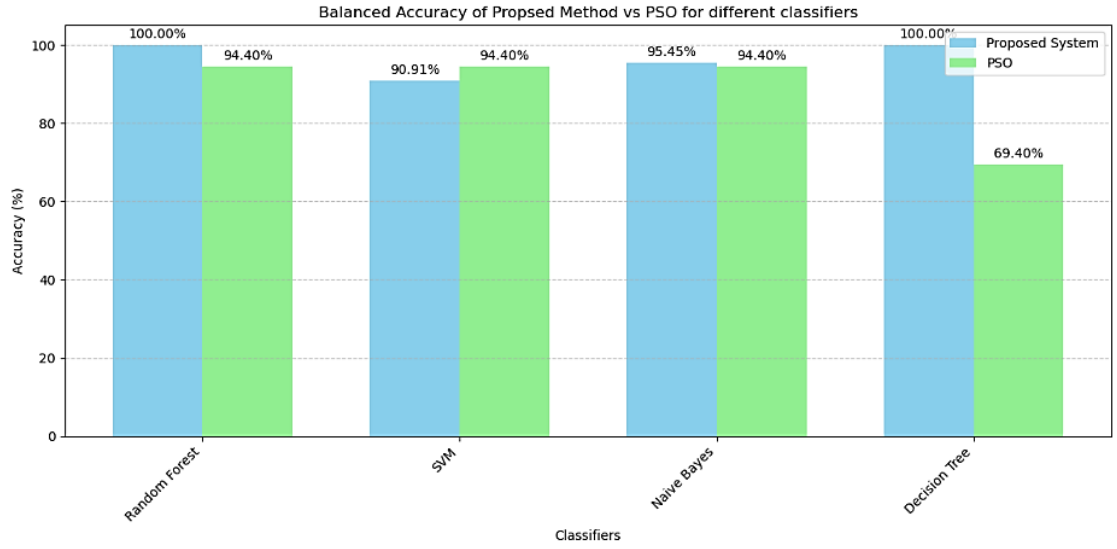

Fig.3. Comparison of Balanced Accuracies of Proposed Method with Particle Swarm Optimization Method

The line chart illustrates the performance of various classification models in terms of test accuracy and balanced Accuracy. Both ensemble models, the Voting Classifier and the Stacking Classifier, achieve the highest possible scores, with 100% test accuracy and 100% balanced accuracy, indicating perfect classification on the test set without class imbalance issues. Similarly, the Random Forest classifier also achieves 100% for both metrics, showcasing its robustness and effectiveness in handling complex gene expression data. The Support Vector Machine (SVM) demonstrates the lowest performance among the evaluated models, with a test accuracy of approximately 87.5% and a balanced accuracy of 90.91%, highlighting its relative difficulty in managing the classification boundaries for this dataset. The K-Nearest Neighbours (KNN) classifier performs moderately well, yielding a test accuracy of 93.75% and a balanced accuracy of 95.45%. Likewise, the Naïve Bayes model exhibits similar results with 93.75% test accuracy and 95.45% balanced accuracy, suggesting consistent performance across different probabilistic classifiers. The Decision Tree classifier matches the top-performing models with a perfect 100% in both test and balanced Accuracy, reinforcing its ability to capture decision boundaries effectively when combined with appropriate preprocessing and tuning. The results validate the superiority of ensemble models and tree-based classifiers, while also revealing that linear models like SVM may underperform without careful kernel and hyperparameter optimization.

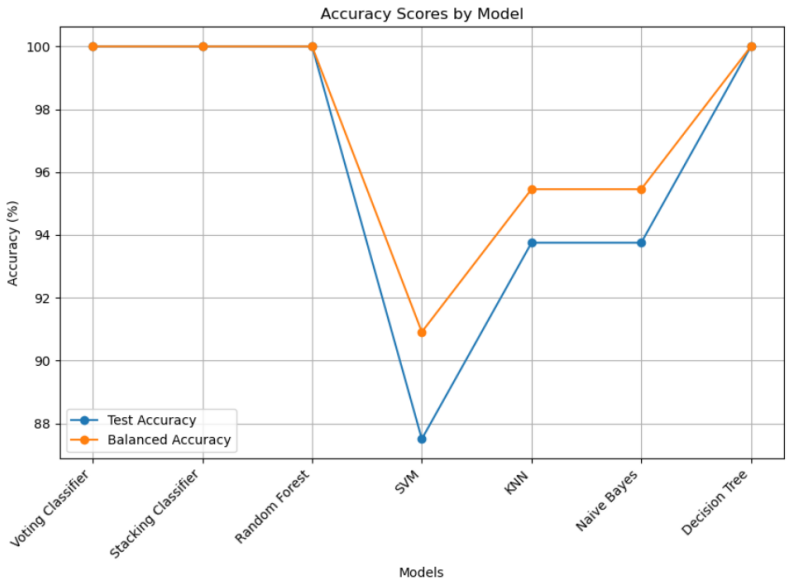

Fig.4. Accuracy Scores of various classifiers

The heatmap presents a comparative visualization of test accuracy and balanced accuracy scores for six different classification models. The Voting Classifier, Stacking Classifier, Random Forest, and Decision Tree each achieve perfect scores of 100% for both test accuracy and balanced Accuracy, indicating excellent model performance without overfitting or class imbalance issues. These models appear in the darkest shade of blue on the heatmap, representing the highest performance. The Support Vector Machine (SVM) shows the weakest results among the models tested, with a test accuracy of 88% and a balanced accuracy of 91%. This noticeable drop, highlighted in a lighter shade on the heatmap, suggests that SVM may struggle with the complexity or distribution of the gene expression data used in this study. The K-Nearest Neighbors (KNN) and Naïve Bayes classifiers both demonstrate strong and consistent performance, each scoring 94% in test accuracy and 95% in balanced Accuracy. These models occupy intermediate color tones in the heatmap, reflecting their solid yet slightly suboptimal performance compared to ensemble and tree-based classifiers. The heatmap visually reinforces that ensemble methods and decision trees are the most effective approaches in this study, while also clearly identifying SVM as the least suited for this specific classification task.

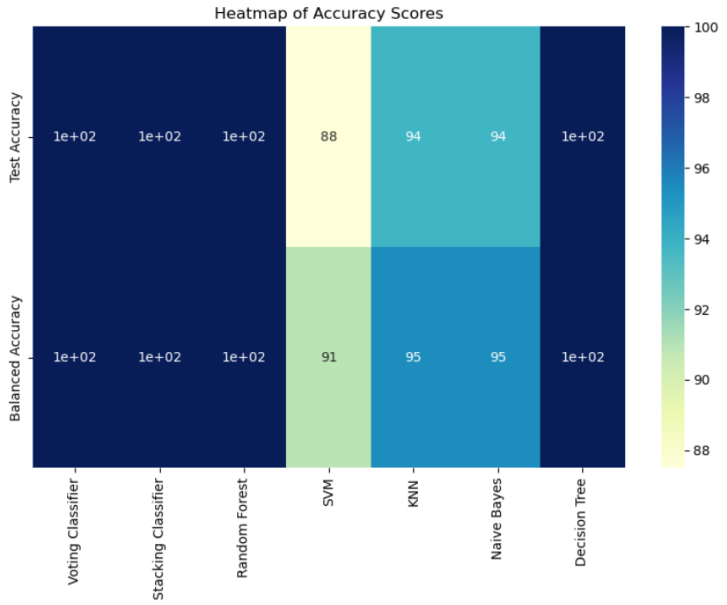

Fig.1. Heatmap of Accuracy Scores of various classifiers

The line chart illustrates the training and validation accuracy of various classification models used in the study. Notably, the Voting, Stacking, Random Forest, and Decision Tree models all achieve a validation accuracy of 100%, indicating that these models are highly effective at generalizing to unseen data. Their corresponding training accuracies range from 92% for the Voting Classifier to 96% for both the Stacking and Random Forest models, and 98% for the Decision Tree, suggesting a balanced fit without significant overfitting. The Support Vector Machine (SVM) model exhibits the weakest performance, with a validation accuracy of just 87.5% and a training accuracy of 96%. This significant performance gap signals potential overfitting, where the model performs well on training data but poorly on unseen validation data. The K-Nearest Neighbors (KNN) model achieves a training accuracy of 98% and a validation accuracy of 93.75%, demonstrating solid generalization capability, albeit slightly below the top-performing ensemble and tree-based models. The chart highlights the superior generalization of ensemble models and Decision Trees, while also revealing that SVM struggles with model stability and adaptability in the context of high-dimensional gene expression data.

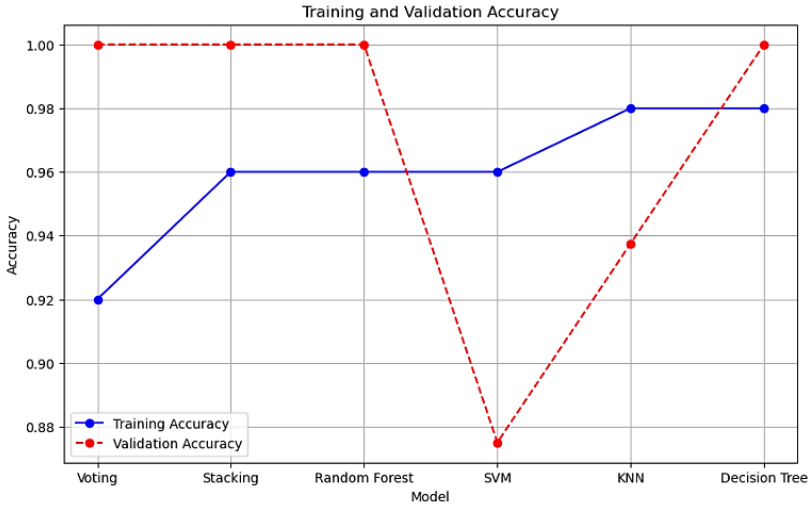

Fig.6. Training and validation Accuracy of various classifiers for 70% Training data and 30% Test data

The collection of learning curves provides insights into the training behavior and generalization performance of five different classifiers: Random Forest, SVM, KNN, Naïve Bayes, and Decision Tree. Each curve plots training score and cross-validation score as a function of the number of training examples, helping to diagnose underfitting, overfitting, and learning dynamics. The Random Forest learning curve shows rapid improvement in cross-validation score as training size increases. Both training and validation scores converge close to 1.0, indicating excellent generalization and minimal overfitting. This suggests that the Random Forest classifier is highly effective for this gene expression dataset, benefiting from ensemble averaging and feature randomness. The SVM curve starts with lower cross-validation performance (~0.78) and improves gradually to around 0.95, while training accuracy remains high (~0.98–1.0). The gap between the curves reduces with more data, implying some initial overfitting that diminishes as more training samples are introduced. The KNN model begins with a training score near 1.0, but the cross-validation score starts lower (~0.85) and increases steadily, reaching parity near 0.95 with more data. This convergence shows that KNN generalizes well as more examples are available, though it initially exhibits overfitting with small data subsets. The Naïve Bayes classifier consistently achieves near-perfect training accuracy (~1.0), but its cross-validation score starts much lower (~0.5) and gradually improves to around 0.90. This consistent gap suggests that Naïve Bayes may be underfitting initially due to its strong independence assumptions but benefits from additional training data. The Decision Tree shows one of the most tremendous improvements, starting with a low validation score (~0.55) and rising sharply to above 0.95. While the training score stays at 1.0, the decreasing gap between the training and validation curves with more examples indicates that the model transitions from overfitting to good generalization. The learning curves affirm that ensemble methods (Random Forest, Decision Tree) and distance-based models (KNN) generalize better with increasing data. At the same time, SVM and Naïve Bayes require careful tuning or larger datasets to close the performance gap between training and validation.

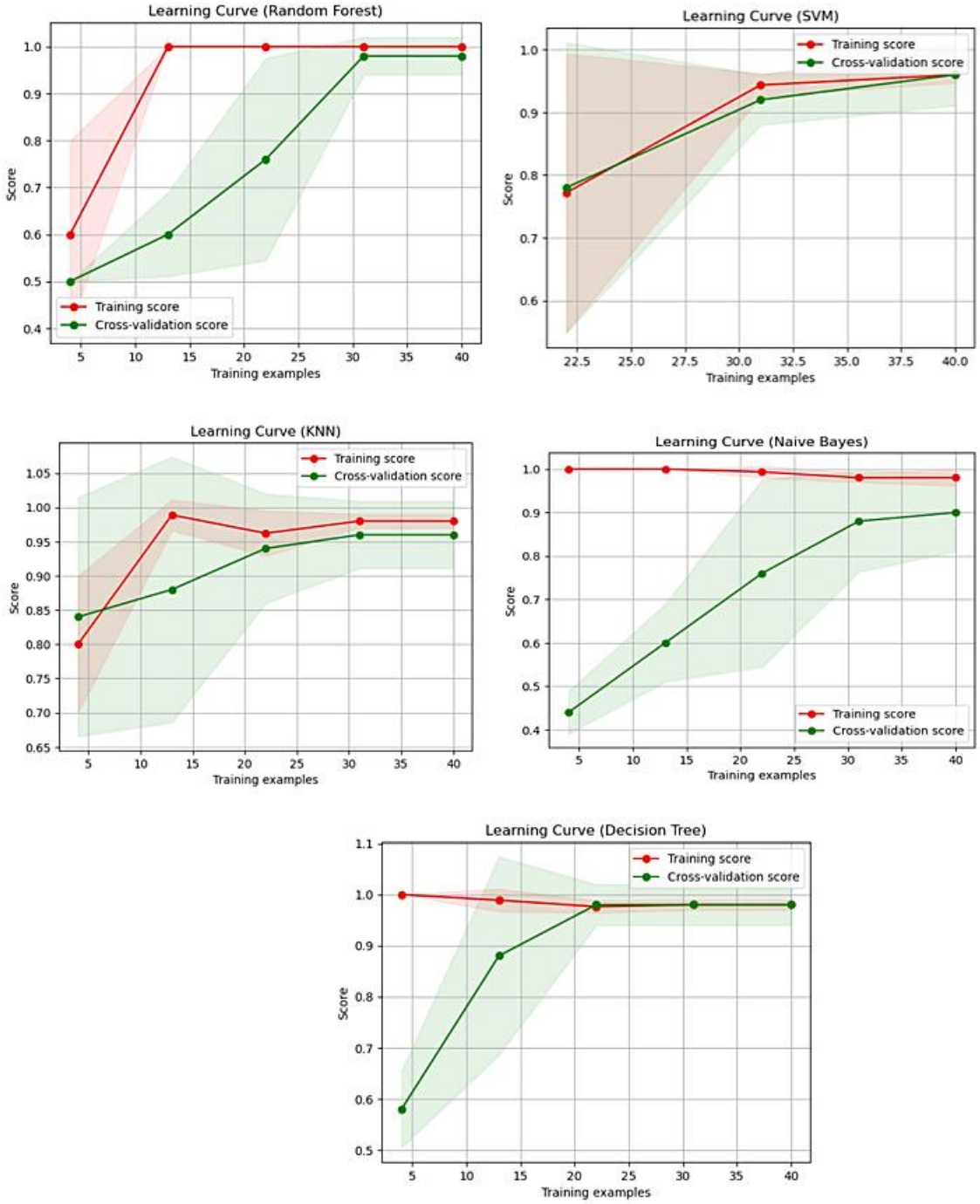

Fig.7. Learning Curve of Various Classifiers

The results clearly demonstrate the superior performance of ensemble models—particularly Random Forest, Voting, and Stacking Classifiers—which consistently achieved perfect or near-perfect scores across multiple evaluation metrics. Decision Tree and KNN also exhibited strong generalization capabilities, while SVM showed comparatively lower performance and signs of overfitting. The proposed methodology, combining robust preprocessing, feature reduction, and ensemble learning, significantly enhances classification accuracy and reliability for high-dimensional gene expression data, making it a promising approach for pancreatic cancer detection.

5. Discussion

The experimental results presented in this study validate the effectiveness of the proposed methodology for classifying pancreatic cancer using microarray gene expression data. Ensemble models such as the Voting Classifier, Stacking Classifier, Random Forest, and Decision Tree achieved perfect scores of 100% for both test accuracy and balanced Accuracy, indicating

exceptional capability in both prediction and class balance handling. These results are consistent across learning curves, heatmaps, and accuracy plots, highlighting not only high model performance but also strong generalization across varying training set sizes. The K-Nearest Neighbors (KNN) and Naïve Bayes classifiers also performed well, with balanced accuracies of 95.45%, confirming their reliability under robust preprocessing. In contrast, the Support Vector Machine (SVM) showed relatively weaker performance, with a test accuracy of 87.5% and balanced Accuracy of 90.91%, coupled with signs of overfitting and higher sensitivity to training data variations.

The learning curve analysis further reinforces these findings. Models like Random Forest and Decision Tree exhibited rapid convergence between training and cross-validation scores, indicating stable learning and minimal overfitting. On the other hand, SVM and Naïve Bayes displayed a larger gap between training and validation scores, suggesting sensitivity to limited training examples and potential underfitting in early stages. These insights confirm that ensemble-based and tree-based models are better suited for the high-dimensional, class-imbalanced nature of gene expression data.

Despite these promising outcomes, the study is not without limitations. First, the dataset used is relatively small, with only 51 total samples. Although methods like SMOTE and cross-validation mitigate this to some extent, larger and more diverse datasets would further validate the scalability and robustness of the proposed approach. Second, the feature selection method used—SelectKBest with ANOVA F-values—assumes linear separability, which may not capture complex interactions among genes. Advanced techniques like recursive feature elimination or deep feature embeddings could provide richer representations. Additionally, model interpretability remains a challenge, especially in ensemble and stacking classifiers, where feature contributions are less transparent. This is particularly important in clinical applications where decisions must be explainable to medical professionals.

For future work, several directions can be explored. Incorporating deep learning architectures such as autoencoders or convolutional neural networks could allow for automatic feature extraction and better capture nonlinear patterns. Multi-omics integration—combining gene expression data with proteomic, metabolomic, or imaging data—could enhance diagnostic Accuracy and biological relevance. Moreover, integrating explainable AI (XAI) frameworks can help bridge the gap between high performance and interpretability, making the models more trustworthy in clinical settings. Finally, external validation using independent datasets and cross-institutional studies would ensure generalizability and clinical applicability of the proposed system.

The proposed ensemble-based methodology provides a robust, accurate, and reliable framework for pancreatic cancer detection using gene expression profiles. While limitations exist, the results mark a substantial improvement over traditional classification approaches and lay a solid foundation for future advancements in precision oncology diagnostics.

6. Conclusion

The study presents a comprehensive and practical methodology for pancreatic cancer classification using high-dimensional microarray gene expression data. By integrating robust preprocessing techniques, feature selection, dimensionality reduction through PCA, and robust ensemble classifiers, the proposed system achieves exceptional performance, with several models—including Random Forest, Voting, and Stacking Classifiers—reaching perfect Accuracy and balanced accuracy scores. Compared to traditional approaches such as those using Particle Swarm Optimization, the proposed framework demonstrates clear improvements in both precision and generalization. While limitations such as dataset size and model interpretability exist, the findings underscore the potential of ensemble learning and advanced preprocessing in improving cancer diagnosis, and pave the way for future exploration involving deep learning, explainable AI, and multi-omics data integration.

Author Contributions

Sreeja Poduri conceptualized the study and designed the robust ensemble learning framework for pancreatic cancer classification using high-dimensional gene expression data. Sreeja Poduri conducted data preprocessing, implemented class imbalance handling using SMOTE, performed feature selection and dimensionality reduction using statistical methods and PCA, and developed and evaluated individual and ensemble machine learning models. Sreeja Poduri analyzed and interpreted the experimental results, validated model performance, and prepared the original manuscript, including revisions and final approval of the submitted version.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] B. Alizadeh Savareh et al., "A machine learning approach identified a diagnostic model for pancreatic cancer through using circulating microRNA signatures," Pancreatology, vol. 20, no. 6, pp. 1195–1204, Sep. 2020, doi: 10.1016/j.pan.2020.07.399.

[2] K. Haseeb, I. Ahmad, I. I. Awan, J. Lloret, and I. Bosch, "A Machine Learning SDN-Enabled Big Data Model for IoMT Systems," Electronics, vol. 10, no. 18, p. 2228, Sep. 2021, doi: 10.3390/electronics10182228.

[3] A. Rghioui, J. Lloret, S. Sendra, and A. Oumnad, "A Smart Architecture for Diabetic Patient Monitoring Using Machine Learning Algorithms," Healthcare, vol. 8, no. 3, p. 348, Sep. 2020, doi: 10.3390/healthcare8030348.

[4] S. P. Menon et al., "An Intelligent Diabetic Patient Tracking System Based on Machine Learning for E-Health Applications," Sensors, vol. 23, no. 6, p. 3004, Mar. 2023, doi: 10.3390/s23063004.

[5] A. Rghioui, A. Naja, J. L. Mauri, and A. Oumnad, "An IoT Based diabetic patient Monitoring System Using Machine Learning and Node MCU," J. Phys.: Conf. Ser., vol. 1743, no. 1, p. 012035, Jan. 2021, doi: 10.1088/1742-6596/1743/1/012035.

[6] B. Kenner et al., "Artificial Intelligence and Early Detection of Pancreatic Cancer: 2020 Summative Review," Pancreas, vol. 50, no. 3, pp. 251–279, Mar. 2021, doi: 10.1097/MPA.0000000000001762.

[7] A. Rghioui, J. Lloret, and A. Oumnad, "Big Data Classification and Internet of Things in Healthcare:," International Journal of E-Health and Medical Communications, vol. 11, no. 2, pp. 20–37, Apr. 2020, doi: 10.4018/IJEHMC.2020040102.

[8] M. R. H. Mondal, S. Bharati, and P. Podder, "Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A Review," CMIR, vol. 17, no. 12, pp. 1403–1418, Dec. 2021, doi: 10.2174/1573405617666210713113439.

[9] K. Haseeb, T. Saba, A. Rehman, I. Ahmed, and J. Lloret, "Efficient data uncertainty management for health industrial internet of things using machine learning," Int J Communication, vol. 34, no. 16, p. e4948, Nov. 2021, doi: 10.1002/dac.4948.

[10] S. Tripathi, A. Tabari, A. Mansur, H. Dabbara, C. P. Bridge, and D. Daye, "From Machine Learning to Patient Outcomes: A Comprehensive Review of AI in Pancreatic Cancer," Diagnostics, vol. 14, no. 2, p. 174, Jan. 2024, doi: 10.3390/diagnostics14020174.

[11] R. Alizadehsani et al., "Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991-2020)," 2020, arXiv. doi: 10.48550/ARXIV.2008.10114.

[12] A. Ogunleye, C. Piyawajanusorn, G. Ghislat, and P. J. Ballester, "Large-Scale Machine Learning Analysis Reveals DNA Methylation and Gene Expression Response Signatures for Gemcitabine-Treated Pancreatic Cancer," Health Data Sci, vol. 4, p. 0108, Jan. 2024, doi: 10.34133/hds.0108.

[13] M. Sinkala, N. Mulder, and D. Martin, "Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics," Sci Rep, vol. 10, no. 1, p. 1212, Jan. 2020, doi: 10.1038/s41598-020-58290-2.

[14] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and Robust Machine Learning for Healthcare: A Survey," 2020, arXiv. doi: 10.48550/ARXIV.2001.08103.