CSCI 5461 Final Project: Machine Learning Methods for Pancreatic Cancer Classification Using Gene Expression Data

Ashley Chen, Vincent Liu, Jerry Yin, Daniel Dowdle & Peiyuan Wei
{chen7361,liu01841,yin00486,dowdl021,wei00334}@umn.edu

Abstract

Pancreatic cancer remains a formidable challenge in medical diagnosis due to its often asymptomatic nature, necessitating advanced classification systems capable of distinguishing between different cancer types based on mRNA features. In this research project, we explore the efficacy of various machine learning models for the classification of pancreatic cancer, focusing on Support Vector Machines (SVM), k-nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF), and Multilayer Perceptron (MLP). Using a dataset of 5 Adenosquamous Carcinoma instances and 135 Pancreatic Ductal Adenocarcinoma (PDAC), our results indicate that SVM and kNN exhibit promising performance on the high-dimensional data, outperforming the dataset imbalance sensitive RF and NB models. MLP, on the other hand, is significantly affected by the small dataset size, exhibiting tendencies towards overfitting. Through experimentation and testing, we present an evaluation of each classic ML model's performance, providing insights into their strengths and limitations in the context of pancreatic cancer classification. Future steps may involve the acquisition of more data, exploration of alternative dataset imbalance strategies, and the implementation of different cross-validation techniques. Our source code can be found in this repo: https://github.com/ac-136/CSCI5461-Final-Project

1 Introduction

Pancreatic cancer stands as one of the most challenging malignancies to diagnose and treat effectively, largely due to its often asymptomatic progression in its early stages. The pancreas, nestled deep within the abdomen, presents unique challenges for detection, resulting in delayed diagnoses and limited treatment options. Among the various types of pancreatic cancer, Pancreatic Ductal Adenocarcinoma (PDAC) and Adenosquamous Carcinoma are predominant, each presenting distinct molecular profiles and clinical behaviors. Distinguishing between these cancer types accurately is crucial for devising tailored treatment strategies and predicting patient outcomes.

The difficulty in classifying pancreatic cancer comes from several factors–the lack of specific symptoms in its initial stages and the heterogeneous nature of pancreatic tumors between patients–resulting in a difficult challenge for accurate classification. Given these challenges, there is a pressing need for advanced classification systems capable of leveraging molecular features to accurately differentiate between various types of pancreatic cancer. mRNA expression profiles are a promising avenue for such classification, and machine learning models, with the ability to analyze high-dimensional datasets, hold immense potential for improving pancreatic cancer classification accuracy and aiding in personalized treatment approaches.

In this research project, we focus on exploring the efficacy of different machine learning algorithms, including Support Vector Machines (SVM), k-nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF), and Multilayer Perceptron (MLP), for the classification of pancreatic cancer subtypes. By utilizing mRNA expression data from PDAC and Adenosquamous Carcinoma instances, we aim to evaluate the performance of these models in accurately distinguishing between different cancer types. Through our investigation, we seek to contribute to the ongoing efforts aimed at

improving pancreatic cancer diagnosis and treatment, ultimately enhancing patient outcomes and survival rates.

2 METHODS

2.1 MODELS

Extensive research has been conducted on cancer classification using a variety of methodologies; however, a significant portion of this work has primarily focused on subclassifications within particular cancer types, often overlooking the potential for broad classification across multiple categories (Alharbi & Vakanski, 2023). This gap in research highlights the need for a comprehensive comparison across both traditional machine learning (ML) and advanced deep learning (DL) algorithms, particularly to understand their efficacy in a more generalized framework. Accordingly, this study aims to critically assess and compare a range of traditional ML techniques alongside advanced DL algorithms, with a focus also on evaluating the impact of various feature selection methods on model performance.

In choosing our models, we have selected four foundational ML algorithms, Support Vector Machines (SVM), k-nearest Neighbors (kNN), Naive Bayes (NB), and Random Forest (RF), due to their diverse mechanisms of operation and broad application in the existing literature. These models represent a spectrum of ML approaches, from SVM's capacity for handling high-dimensional data and kNN's efficacy in classification via proximity, to NB's probabilistic approach and RF's ensemble method that enhances prediction accuracy and overfitting control. These characteristics make them ideal benchmarks in the comparative analysis of cancer classification methodologies.

Against these traditional models, we position a Multilayer Perceptron (MLP), a fundamental structure in deep learning known for its ability to learn non-linear relationships through layers of interconnected nodes (Khalid et al., 2014). The inclusion of MLP allows us to explore the adaptability and strengths of deep learning in capturing complex patterns that may not be as readily discerned by traditional ML algorithms.

SVM. Support Vector Machine (SVM) is a powerful and versatile supervised machine learning algorithm used primarily for classification, though it can also be adapted for regression tasks. The main idea behind SVM is to find a hyperplane in an N-dimensional space (N being the number of features) that distinctly classifies the data points into different categories. They are particularly well-suited for complex but small or medium-sized datasets and can be very effective even when the data has a clear margin of separation. They are less effective on very large datasets or datasets with lots of noise, where the class overlap significantly reduces the clarity of the margin. SVM are favored for cancer classification due to their robustness in high-dimensional spaces, such as gene expression data, effectively managing complex and non-linear relationships through the kernel trick and minimizing overfitting with regularization. This allows SVMs to maintain clear decision boundaries and reliable generalization, crucial for accurate medical diagnostics.

kNN. k-Nearest Neighbors (kNN) is a simple and intuitive supervised machine learning algorithm used for both classification and regression tasks. It operates on a very straightforward principle: to predict the class of a given data point, kNN identifies the k closest points in the training dataset, known as neighbors, and determines the output based on their classes or values. The majority class among these neighbors decides the class for classification tasks, while the average or median of their values is used for regression. It is a non-parametric method, meaning it makes no assumptions about the underlying data distribution. This feature, combined with its ease of implementation, makes kNN a popular choice for many introductory machine learning problems, especially when a quick and effective solution is needed. It is chosen for cancer classification because of its simplicity and effectiveness in handling small, well-curated datasets where relationships between features and class labels are straightforward. It excels in scenarios where the proximity of similar cases can strongly predict outcomes, leveraging the majority rule among nearest neighbors for robust, interpretable classifications.

NB. Naive Bayes (NB) is a simple yet powerful probabilistic machine learning model based on Bayes' Theorem, used predominantly for classification tasks. It operates under the "naive" assumption that features in the dataset are mutually independent given the class label. Its efficiency, ease of

implementation, and good performance with large datasets and high-dimensional spaces make it a popular choice for many introductory machine learning problems. It is chosen for cancer classification due to its efficiency and effectiveness in handling large datasets with multiple features, such as genetic markers, by assuming feature independence and leveraging probabilistic inference. Its ability to quickly model and predict class probabilities based on feature presence makes it particularly useful for initial screening and diagnostic tests in medical settings.

RF. Random Forest (RF) is an ensemble machine learning algorithm that builds on the simplicity of decision trees by creating a "forest" of them to improve predictive accuracy and control overfitting. It operates by constructing multiple decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random Forests are effective because they reduce the variance of individual trees through averaging and are less likely to overfit than a single decision tree. They handle both numerical and categorical data well, can model complex interactions between features, and are robust against noise, making them versatile for a wide range of applications. It is selected for cancer classification because it can handle complex and high-dimensional datasets, like those common in medical diagnostics, by utilizing multiple decision trees to enhance prediction accuracy and prevent overfitting. This method effectively captures intricate patterns in data, which is crucial for reliable identification and classification of various cancer types.

MLP. Multilayer Perceptron (MLP) is a type of artificial neural network that is widely used in deep learning for solving complex pattern recognition and classification problems. An MLP consists of at least three layers of nodes: an input layer, one or more hidden layers, and an output layer. Each node, or neuron, in one layer connects with a certain weight to every node in the following layer, making the network fully connected. MLPs use a method known as backpropagation for training, they learn by iteratively adjusting the weights of connections to minimize the difference between the actual output and the predicted output. It is chosen for cancer classification because of its ability to model complex, non-linear relationships inherent in biological data, despite being prone to overfitting when trained on small datasets. Techniques such as regularization, dropout, and careful validation can mitigate overfitting, making MLPs effective for capturing subtle patterns critical for accurate cancer diagnosis.

3 RESULTS

3.1 RESULTS PRE-DATASET IMBALANCE CORRECTION

Model	TN	FP	TP	FN
SVM	0	5	135	0
KNN	0	5	135	0
NB	0	5	135	0
RF	0	5	135	0
MLP	0	5	135	0

Table 1: Model results on initial dataset

The dataset we used was quite unbalanced with 135 samples of PDAC and only 5 Adenosquamous carcinoma. We used Leave-One-Out (LOO) cross validation in response to the samll data size and severe imbalance. However, even using LOO cross validation was not enough to overcome these dataset issues. These dataset issues led the results to be extremely skewed. The results would correctly predict PDAC, but due to the lack of samples of Adenosquamous, they would incorrectly classify them consistently. In other words, regardless of the RNA sequence, the model would always predict PDAC. This is shown in Figure 1. After receiving these results we pivoted to fixing the dataset imbalance in order to better results. After fixing the dataset imbalance, the results were much more diverse and promising for a few of the models.

3.2 RESULTS POST DATASET IMBALANCE CORRECTION

We tried multiple approaches to combat the dataset imbalance. We originally tried oversampling. Oversampling is a common data imbalance correction technique in which random duplicates are

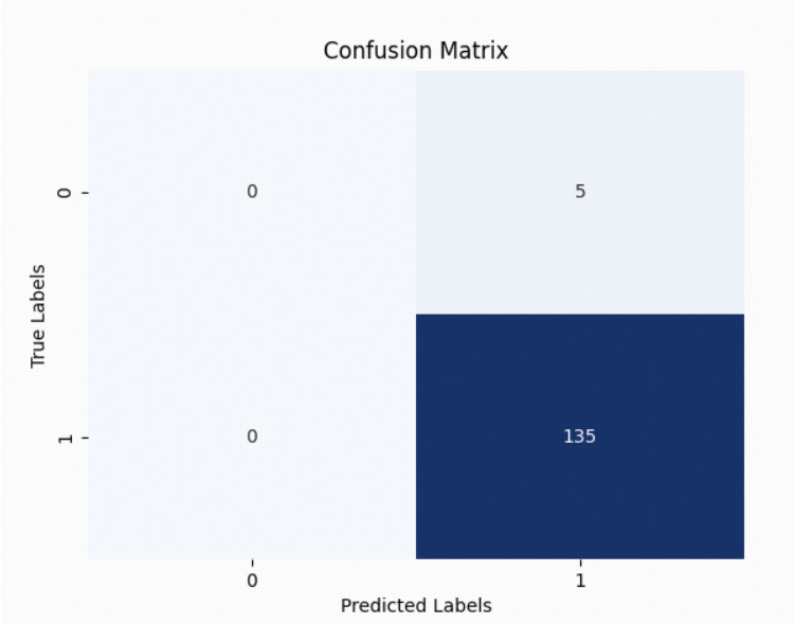

Figure 1: Confusion matrix for all models tested on original dataset using Leave-One-Out cross validation. Dataset imbalance caused overclassification of of PDAC (class 1).

made of the minority class. Unfortunately, oversampling was not very effective. We believe this is because of the small number of Adenosquamous carcinoma. We tried using different ratios of Adenosquamous carcinoma to PDAC, but regardless, we were unable to stop the overclassification of PDAC.

Since oversampling didn't work, we tried undersampling. Instead of duplicating samples of the minority class, undersampling randomly removes samples from the majority class. We started to see better results. Some of the models were able to correctly predict Adenosquamous carcinoma, but it was only one or two correct predictions for only a couple of the models.

Therefore, we tried the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE similar to oversampling, increases the amount of samples of the minority class. However, instead of duplicating existing samples, SMOTE synthesizes new samples based on the existing minority class. SMOTE works by creating a feature space based on the k-nearest neighbors of a sample of the minority class. This feature space is then used to generate a new sample of the minority class. SMOTE, unfortunately, was unable to correct the dataset imbalance. We believe this is because there are not enough samples of the minority class. Attempting to create over 100 new samples to better the ratio between the minority and majority class, from just five samples in simply infeasible.

However, we still believed we could improve our model performance. Given the promising results of undersampling, we decided to combine undersampling with SMOTE. We were finally able to get positive results and significant differences between models. Our results are shown in Table 2 and our two best model results are displayed in a confusion matrix in Figure 2. Overall, SVM and KNN had the best results. Unlike other models, these two models were able to get over 80% accuracy for both the Adenosquamous carcinoma class and the PDAC class. Both models only misclassified one Adenosquamous carcinoma sample. Additionally, both models were able to correctly classify around 110 PDAC samples.

4 Discussion

Based on the experiment results, we tried to analyze the reason for the performance of the models and guidance to future studies in this area.

Model	TN	FP	TP	FN
SVM	4	1	112	23
KNN	4	1	110	25
NB	4	1	41	94
RF	3	2	83	52
MLP	0	5	133	2

Table 2: Model results on revised dataset using undersampling and SMOTE.

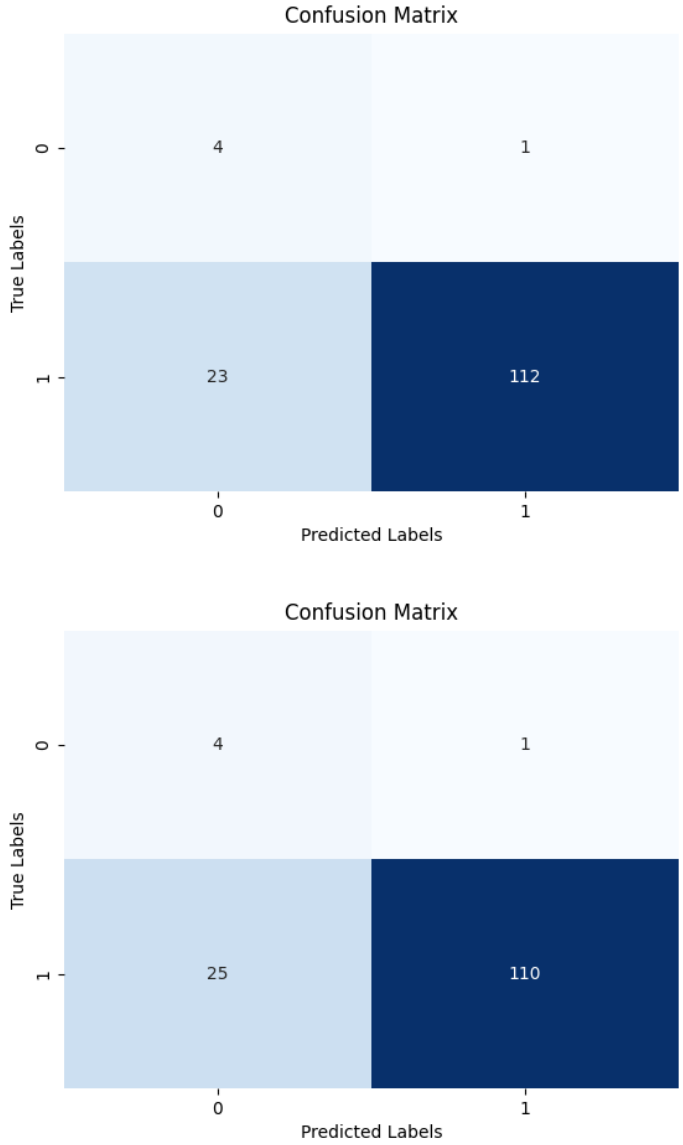

Figure 2: Our two best model results using undersampling and SMOTE: SVM (top) and KNN (bottom). Only these two models were able to get 80% or higher accuracy for both classes (Adenosquamous carcinoma = class 0, PDAC = class 1)

4.1 Analysis of Models

SVM. The support vector machine is able to construct a hyperplane in a high-dimensional space, and genomic data is a kind of high-dimensional data. Therefore, it is reasonable that SVM works better on this dataset. What is more, in our experiments, when SVM uses a linear kernel, it performs no better than other models, but when it uses a non-linear kernel, such as a polynomial kernel, it performs better. What's more, it works better when the degree of polynomial kernel is 3 than when it is 2 or 4. This suggests that the hyperplane that classifies in this dataset might also be non-linear and this task is possibly a non-linear cubic classification task. This conclusion is important in following an analysis of other models.

kNN. The k-nearest neighbors algorithm gives predictions based on k-nearest neighbors, therefore it also performs well on high-dimensional data. This is a possible reason that it works well on this dataset. Also, k-nearest neighbors is a non-linear classier, thus it will work on this non-linear classification.

Native Bayes. In our experiments, Native Bayes has so many false negative predictions that are even more than true positive predictions. Therefore, it is reasonable to guess Native Bayes is underfitting in this task and does not grab enough features from the given data. Considering SVM works best when it has a polynomial kernel with the degree at 3, and Native Bayes is usually considered a linear or quadratic classifier, underfitting probably results from its inability to cubic classification.

Random Forest. In our experiments, Random Forest does not perform well. As known to all, Random Forest is based on the decision tree, and the decision tree is a technique that is especially sensitive to dataset imbalance. Although we conducted a combination of undersampling and SMOTE to balance the dataset, it is still not perfect and can possibly have a pretty negative influence on Random Forest, which contributes to its' bad performance.

MLP. In our experiments, Multilayer Perceptron gives 135 true positive predictions among 140 predictions, which is far better than all other models and seems extremely perfect. However, as we have to use the training dataset as a testing dataset due to the small size of data we have, it is possible that MLP is just overfitting in this very small dataset because it does not give any true negative predictions. Although we only used 1 layer in our MLP models, it is still a relatively complex model and thus is likely to overfit the PDAC class.

4.2 Future Steps

We are examining the next steps for our research on pancreatic cancer classification using machine learning. As our initial work highlights the need for better data handling and model validation to improve accuracy and reliability, future studies can focus on the following topics:

- Given the complexity and rarity of cancer, large datasets are crucial to develop robust models. Aiming to boost prediction precision and the model's generalization across diverse patient groups and cancer subtypes, we have to enlarge our dataset. This can also avoid overfitting from some advanced models.

- In our experiments, we learned that it's essential to address dataset imbalance. In this project, we utilized and evaluated several methods including oversampling, undersampling, and SMOTE. In future studies, more advanced techniques such as adaptive synthetic sampling and borderline-SMOTE, can be applied.

- In this project, we do not employ deep learning methods on this small dataset. To get better results, in future studies, we can test more advance deep learning models on larger datasets.

- In this project, we used leave one out cross-validation to combat the small size of the dataset. In the future, We also aim to explore various cross-validation strategies, including stratified k-fold and time-series cross-validation, tailored to our data collection methods on larger datasets.

REFERENCES

Fadi Alharbi and Aleksandar Vakanski. Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, 10:173, 01 2023. doi: 10.3390/bioengineering10020173.

Samina Khalid, Tehmina Khalil, and Shamila Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pp. 372–378, 2014. doi: 10.1109/SAI.2014.6918213.