ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
Department of Computer Science and Engineering (CSE)
A Subsidiary organ of the organization of Islamic Cooperation (OIC)
Dhaka, Bangladesh

Predicting cancer origins with DNA Methylation

Authors' name and IDs

Chaanraoui Ben Djoumoi	180041156
Charif Abdallah Yahaya Charif	180041253
Rabianti Said Youssouf	180041254

Supervisor

Tareque Mohmud Chowdhury

Assistant Professor
Department of Computer Science and Engineering

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of Bachelor of Science in Computer Science and Engineering (CSE)

May 2023

DECLARATION

This is to confirm that the students indicated below worked under the supervision of Mr. Tareque Mohmud Chaowdhury, Assistant Professor at the Department of Computer Science and Engineering at the Islamic University of Technology (IUT). This article represents the culmination of the student's thesis work for the Bachelor of Engineering in Computer Science degree.

.....

Chaanraoui Ben Djoumoi *(180041156)*

.....

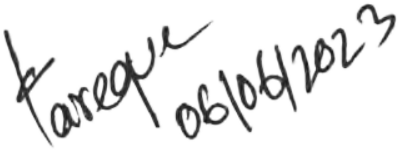
Charif Abdallah Yahaya Charif*(180041253)*

.....

Rabianti Said Youssouf*(180041254)*

CERTIFICATE OF RESEARCH

Chaanraoui Ben Djoumoi (180041156), Charif Abdallah Yahaya Charif (180041253), and Rabianti Said Youssouf (180041254) have submitted a thesis titled "PREDICTING CANCER ORIGINS WITH DNA METHYLATION" a prerequisite for the Bachelor of Science in Computer Science and Engineering degree has been acknowledged as partially satisfied in a suitable manner.



.....

Tareque Mohmud Chowdhury

Assistant Professor

Department of Computer Science and Engineering

ACKNOWLEDGEMENT

We would like to thank God Almighty for leading us through all of our difficulties. We have sensed your leading us day by day. You are the one who made it possible for us to earn our degree. We'll keep having trust in you for our future. We also like to thank our supervisor Tareque Mohmud Chowdhury, who made it possible for us to finish this work. His advice and direction helped us at every level of the essay-writing process. We also want to express our gratitude to the committee members for their thoughtful remarks and recommendations, which made the experience of giving our defense enjoyable. We would want to express our gratitude to our parents and families as a whole, as well as to the Almighty, for their spiritual support, never-ending prayers, and inspiration they provided over the course of our study. We are eternally appreciative of our family and friends for their assistance.

Contents

Chapter 1

1.1 Introduction

Predicting the origin of cancer is an important step in cancer diagnosis and treatment planning. Accurately identifying the type of cancer and its origin can help doctors determine the most effective treatment strategy and improve patient outcomes. DNA methylation patterns have emerged as a potential biomarker for predicting the tissue of origin in various types of cancer. DNA methylation is an epigenetic modification that regulates gene expression and plays a critical role in cellular differentiation and development. Aberrant DNA methylation patterns have been observed in various cancers, and specific methylation signatures have been associated with different cancer types. Analyzing DNA methylation profiles can provide valuable insights into the tissue of origin of cancer and help guide clinical decision-making.

DNA methylation is an epigenetic modification that involves the addition of a methyl group to the cytosine base of DNA, typically at CpG dinucleotide sites. It can regulate gene expression by blocking the binding of transcription factors or recruiting proteins that modify chromatin structure, leading to changes in gene expression patterns. Abnormal DNA methylation patterns have been observed in various types of cancer, and these patterns can serve as potential biomarkers for cancer detection, diagnosis, and prognosis. DNA methylation profiling has become a powerful tool for predicting cancer's origin. By analyzing DNA methylation patterns in cancer cells, researchers can identify specific methylation signatures associated with different tissue types, which can be used to develop predictive models that accurately classify the tissue of origin of cancer based on DNA methylation data. These models can have significant implications for cancer diagnosis, treatment, and patient management.

Random Forest and K-Nearest Neighbors (KNN) are two popular machine learning algorithms used in predicting cancer origin based on DNA methylation data. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. K-Nearest Neighbors is a proximity-based classification algorithm that assigns new samples to the class of their K nearest neighbors based on a distance metric. A comparative analysis of these algorithms can shed light on their effectiveness in predicting the origin of cancer and provide insights into the factors influencing the choice of algorithm for different scenarios. Both Random Forest and KNN algorithms have their strengths and limitations, and comparative analysis can shed light on their effectiveness in predicting the origin of cancer and provide insights into the factors influencing the choice of algorithm for different scenarios.

1.2 Problem Statement

Researchers and clinicians are working to improve existing techniques for predicting the origin of cancer-based on pathology and gene expression. Machine learning and AI algorithms are being used to analyze large-scale genomic and transcriptomic data, single-cell analysis and multi-omic integration approaches are being used to understand intra-tumor heterogeneity and liquid biopsies are being explored to analyze tumor-derived material from blood samples. Collaborative efforts among researchers, clinicians, and data scientists are essential to addressing these challenges and improving cancer origin prediction.

1.3 Objectives

Cancer is a dangerous disease that affects people worldwide, and its occurrence has been increasing rapidly in recent years. With the rapid development of computer science and machine learning technology, computer-aided cancer prediction has significantly improved. DNA methylation, an important biological process, plays a significant role in cancer's origin and progression, making it a potential marker for cancer identification. In our research, we aim to use machine learning algorithms, such as Random Forest and K-Nearest Neighbors (KNN), to predict cancer origins by analyzing DNA methylation data.

Chapter 2

Literature review

2.1 paper 1: Predicting cancer origins with a DNA methylation-based deep neural network model [ZX20]

Zheng C and Xu R published a paper on a high-performance and computation-efficient deep neural network method for predicting cancer origins with DNA methylation data of 7339 patients of 18 different cancer origins from The Cancer Genome Atlas (TCGA). This DNN model produced predictions for various cancer tissue origins with an accuracy of 100 % and an average AUC of 0.99. This model showed high performance in predicting the cancer tissue origins of solid tumors and is also used to identify cancer cell types such as CTCs (circulating tumor cells) and CUPs (cancers of unknown primary). Our DNN model demonstrated greater accuracy (95.03 % vs. 89.4 %), recall (92.3 % vs. 87.8 %), and specificity (99.7 % vs. 99.4 %) when compared with Pathwork, a commercially available cancer origin classifier based on gene expressions.

2.2 Paper 2: Sarcoma classification by DNA methylation profiling [Koe+21]

Christian Koelsche, Daniel Schrimpf, and Damian Stichel et al. suggested t-Distributed Stochastic Neighbour Embedding and unsupervised hierarchical clustering as methods for analyzing methylation data in order to find tumor groupings that share methylation patterns. A dataset of 1077 methylation profiles from cases that were well pre-characterized was used to train this sarcoma classifier. This dataset included 62 tumor

methylation classes that represented a wide variety of soft tissue and bone sarcoma subtypes throughout the full age spectrum. DNA methylation data can be used to categorize sarcomas, although its accuracy is limited. On the other hand, the research also noted that Ewing sarcoma could be distinguished from other cell types with nearly 100 % accuracy. When tumor cells make up 70 % or more of a sample's total cells, our expertise is at its finest. [Koe+21]

2.3 Paper 3: DNA Methylation Markers for Pan-Cancer Prediction by Deep Learning [Liu+19]

Deep-learning DNA methylation biomarkers for pan-cancer have been reported by Biao Liu, Yulu Liu, Xingxin Pan, Mengyao Li, Shuang Yang, and Shuai Cheng Li. They made their work publicly available on October 4, 2019, at PubMed. They used machine learning to discover DNA methylation signatures, and they built diagnostic prediction models with deep learning. They discovered two kinds of markers: 12 CpG markers and 13 promoter markers. Three of the twelve CpG markers and four of the thirteen promoter markers identify cancer-related genes. Their model employing CpG markers exhibited an average sensitivity and specificity of 92.8 % and 90.1 % on test data sets, respectively. The average sensitivity and specificity for promoter markers in test data sets were 89.8 % and 81.1 %, respectively. Furthermore, in cell-free DNA methylation data from 163 prostate cancer patients, CpG markers had a sensitivity of 100 %, whereas promoter markers had a sensitivity of 92 %. For both marker categories, normal whole blood had 100 % specificity. Finally, they discovered methylation markers that can be utilized to diagnose pan-cancers and may be employed in cancer liquid biopsies.

2.4 Paper 4: Minimalist approaches to cancer tissue-of-origin classification by DNA methylation [Xia+20]

A Canadian Academy of Pathology 2020 (2020) research team led by Daniel Xia investigated minimalist methods for cancer tissue-of-origin categorization by DNA methylation. Using data from only 53 of the approximately 450,000 available CpG probes, one classifier from The Cancer Genome Atlas Network achieved an accuracy of 94.5 % on 2575 brand-new primary validation cases across 28 cancer types. The most useful

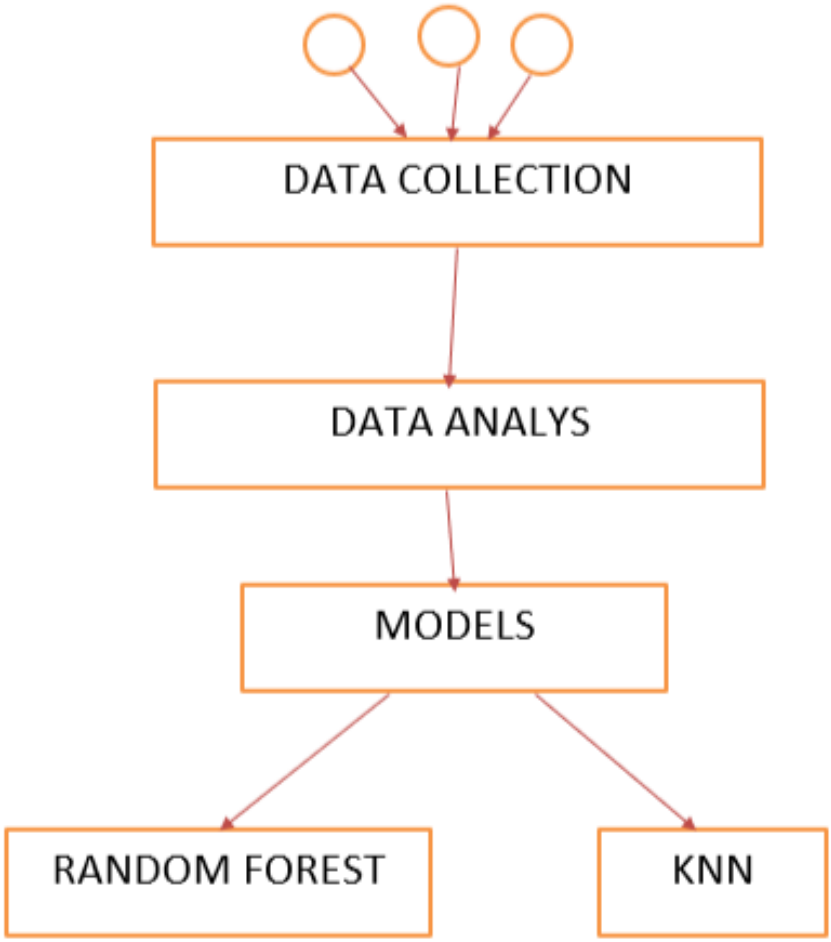
diagnostic probes were identified using DNA methylation profiling data from the Cancer Genome Atlas (TCGA). There were 2188 distinct CpG sites, which corresponded to 1176 distinct genes. They discovered that the genes were enriched for various biological processes, such as tissue formation and morphogenesis, using Enrichr.

Chapter 3

Research methodologies

3.1 Research methodologie

To address the machine learning issue and deploy our approach, we have emphasized four critical steps. To ensure the efficacy and efficiency of our strategy, these steps have been carefully selected. We are primarily focused on simplifying the development process and obtaining exact results throughout these stages. The model's overall performance is impacted differently by each phase. We carefully examine the pre-processing methods available and determine which ones will best prepare the dataset for modeling. Then, to improve the predictive power of our models, we use appropriate techniques for feature selection. Then, to maximize performance, we start the model training and assessment phase. For this, we make use of reliable algorithms and validation procedures. We believe that by adhering to these four crucial steps, we will be able to create a comprehensive machine-learning solution that successfully solves the issue at hand.



3.2 Data Collection and Description

We have used a dataset that has been collected from a repository on GitHub. The data provided are from GEO, which contains 10 cancer origins and 581 patients. We can access the repository from the link below: Access the GitHub repository.

The dataset contains 580 rows and 10360 columns. They found those data from TCGA.

	cg00003994	cg00005847	cg00008493	cg00008713	cg00015770	cg00016968	cg00019495	cg00022866	cg00024396	cg00027083
0	0.222203	0.556290	0.861410	0.139675	0.297772	0.632403	0.311314	0.596050	0.155746	0.242072
1	0.139818	0.630202	0.866165	0.156778	0.220811	0.587911	0.176913	0.743849	0.159064	0.117798
2	0.128091	0.683754	0.868291	0.130763	0.591848	0.533164	0.249263	0.724940	0.112388	0.161069
3	0.583029	0.857389	0.856419	0.140898	0.733095	0.788900	0.692803	0.794868	0.116704	0.033696
4	0.146785	0.872714	0.875467	0.144017	0.161092	0.550712	0.158803	0.490997	0.143484	0.067864

5 rows × 10361 columns

Figure 3.1: the first five rows of the dataset used

cg27641018	cg27643859	cg27644292	cg27648946	cg27650434	cg27651218	cg27652350	cg27653134	cg27661264	primary_code
0.577732	0.821495	0.536977	0.294026	0.143827	0.884275	0.251016	0.680305	0.452540	0
0.736741	0.882523	0.447426	0.195170	0.130605	0.689605	0.273031	0.718857	0.361774	0
0.674942	0.891105	0.364653	0.312903	0.131209	0.867614	0.638708	0.839373	0.443303	0
0.848902	0.888610	0.518056	0.152965	0.297748	0.904576	0.746427	0.878620	0.625918	0
0.707924	0.864790	0.280180	0.164823	0.121440	0.887591	0.170775	0.622952	0.478879	0

Figure 3.2: the first five rows of the dataset used

3.3 Data Analysis

Before applying machine learning algorithms, it is crucial to thoroughly understand and analyze the dataset. This involves conducting statistical analysis and visualizing the data distribution for different variables. We have confirmed that the dataset provided by the supplier has already been cleansed and does not contain any missing values, so there is no need to delete any columns or rows. The dataset has a size of (5580, 10361), indicating 5,580 rows and 10,361 columns. However, the high number of columns suggests a potentially large number of features, which can present challenges in analysis and modeling. Therefore, it is important to carefully explore the data. Descriptive statistics can provide insights into numerical variables, while visualization techniques such as histograms, box plots, scatter plots, and bar plots can help understand the distribution and relationships between variables. Additionally, correlation analysis can uncover potential dependencies among variables. If the dimensionality of the dataset is too high, dimensionality reduction techniques like PCA or t-SNE can be applied. By following these steps, you can gain a deeper understanding of your dataset and make informed decisions when applying machine learning algorithms.

3.4 Models

We must now construct the models after improving the data's quality, normalizing the data, and choosing the organization. In this study, we only created two classifiers. We divided the data into training and testing throughout the training phase. We used 60 % of the data for training and 40 % for testing. Finally, using test data, we trained and evaluated our models. Let's first comprehend how each classifier makes predictions in the background before creating the models.

3.4.1 Random Forest and Decision Tree

Random Forest is a well-known supervised learning technique that may solve classification and regression issues. To solve increasingly difficult problems efficiently, it employs an ensemble learning approach in which numerous classifiers are merged into a single classifier. During the training phase, Random Forest generates many internal decision trees, each with its own output. Each decision tree predicts a class label for classification and a numerical value for regression. Random Forest integrates the results of all decision trees to provide the final forecast. In classification, the final class label is determined using a majority vote process based on the most widely anticipated class. The final prediction in regression can be produced by averaging or taking the median of the projected values. Random Forest integrates the results of all decision trees to provide the final prediction. In classification, the final class label is determined using a majority vote process based on the most widely anticipated class. The final prediction in regression can be determined by averaging or taking the median of the anticipated values from all decision trees. This ensemble technique improves Random Forest's resilience and accuracy, lowering the danger of overfitting and boosting generalization performance. Random Forest is an excellent choice for dealing with high-dimensional data, accepting missing values, and capturing complicated correlations between variables. In the end, it is a flexible and powerful algorithm that generates solid predictions by aggregating the results of its constituent decision trees.

Features of a Random Forest Algorithm

- Compared to the decision tree algorithm, it is more accurate.
- It offers a practical method for dealing with missing data.
- Without hyper-parameter adjustment, it can generate a respectable forecast.

- It fixes the over-fitting problem with decision trees.
- At the node's splitting point in every random forest tree, a subset of characteristics is chosen at random.

Understanding decision trees

Decision trees are a prominent algorithm for predicting cancer, with their intuitive structure and ability to handle both categorical and numerical data. They operate by partitioning the data into subsets based on feature values and recursively constructing a tree-like structure that represents decision rules. They are interpretable, providing clear decision paths that can be easily understood by medical professionals. They are also capable of handling both categorical and numerical features, making them suitable for diverse cancer prediction datasets. Categorical features, such as genetic mutations or histological characteristics, can be encoded as binary variables or using more sophisticated techniques like one-hot encoding.

Decision trees are an advanced cancer prediction technique that provides interpretability and the ability to handle a wide range of data types. They do, however, have disadvantages, including overfitting the training data and difficulties identifying complicated correlations between elements. Ensemble approaches like random forests or gradient boosting can be used to increase prediction accuracy. However, decision trees are not the only option for cancer prediction, and the model used is determined by the problem's complexity and the available resources. Researchers and healthcare practitioners may progress cancer prediction and improve patient outcomes by using the capabilities of decision trees and investigating other models when appropriate.

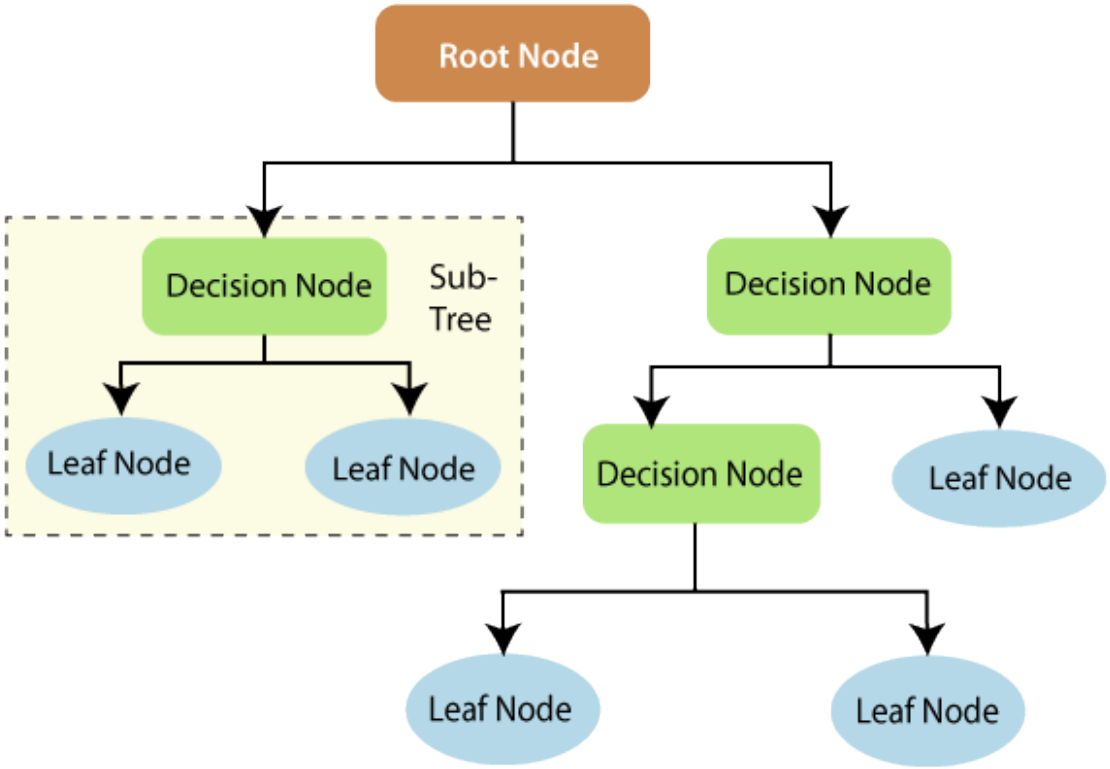


Figure 3.3: Sorting with random forests

Classification in random forests

Random forest classification achieves the result using an ensemble process. Different decision trees are trained using the training data. This dataset is made up of observations and characteristics that will be chosen at random when the nodes are being separated. Different decision trees are used in a rainforest system. Each decision tree has a root node, leaf node, and decision node. The leaf node of each tree represents the result generated by a particular decision tree. The majority voting method is used to choose the final product. In this scenario, the final output of the rainforest system is the output that the majority of the decision trees have selected. A straightforward random forest classifier is depicted in the diagram below.

The use of a decision tree

- The first stage is to collect a dataset with relevant characteristics and labels indicating the presence or absence of cancer.
- It is critical to find the most useful characteristics in order to develop an accurate decision tree model. This is accomplished by the use of feature selection approaches such as statistical testing, correlation analysis, or domain expertise. Choosing the appropriate characteristics improves the model's performance and interpretability.
- The decision tree method constructs a tree-like structure from the acquired data. The method iteratively separates the data depending on the specified criteria, with the goal of maximizing cancer case similarity within each branch or leaf of the tree. The splitting process continues until a requirement, such as reaching a predetermined depth or a certain amount of samples in a leaf, is reached.
- Once built, the decision tree can be used to forecast whether a patient is likely to have cancer or not. Given a set of feature values for a patient, the tree traversal procedure starts at the root node and directs the patient along the proper branches based on the feature values until it reaches a leaf node. The forecast at the leaf node corresponds to the cancer outcome prognosis.
- The decision tree model's accuracy and performance must be evaluated using evaluation measures such as accuracy, precision, or recall. This helps assess the model's accuracy in predicting cancer and can lead to future improvements or modifications.

- The interpretability of decision trees is one of its advantages. The decision tree's structure facilitates simple comprehension of the prediction process. The decision criteria at each node give insights into the relevant traits and thresholds, allowing physicians and researchers to understand the aspects impacting cancer prediction.
- Using combined methods such as random forests, decision trees can be improved further. Multiple decision trees are used in these strategies to increase accuracy and prevent overfitting. Furthermore, hyperparameter tweaking may be used to improve model performance by modifying parameters like maximum tree depth, minimum sample split, or impurity measurements.

3.4.2 K-Nearest Neighbors

KNN is a supervised learning strategy, like Random Forest and Decision Tree. By computing the distance between a data point and all the training data, KNN tries to predict the label of a data point. Useful distance measures include the Manhattan Distance and Euclidean Distance.

Why did we choose K-nearest Neighbor ?

KNN is a non-parametric method, which means it makes no assumptions about the data distribution. DNA methylation patterns can be exceedingly complicated and non-linear, and KNN can successfully capture these patterns without making any assumptions. It works by locating the K closest neighbors to a given data point and categorizing it based on the majority class among those neighbors. This can be useful in the context of cancer prediction using DNA methylation since surrounding samples are likely to have comparable methylation patterns if they come from the same tissue or origin. As a result, KNN may detect localized decision boundaries that correspond to underlying biological traits. Noise in DNA methylation data can occur owing to a variety of circumstances, including experimental variability. Because it considers several neighbors for classification, KNN is particularly resistant to noisy data. Outliers or noisy samples are less likely to dominate the decision-making process, minimizing noise's influence on forecasts. KNN produces outcomes that are easy to understand. Because it classifies a sample based on the majority class of its nearest neighbors, it is possible to explain the prediction by referring to those nearby samples. This interpretability is useful in biological research because it enables researchers to obtain insights about sample similarities and comprehend the basis for their predictions.

How does K-Nearest Neighbor work?

The KNeighborsClassifier, which implements the KNN algorithm, has been imported. In addition, the

classification_report function is imported to provide a classification metrics report. The KNeighborsClassifier class is instantiated. The n_neighbors option is set to 20, indicating that the algorithm will classify the algorithm's 20 closest neighbors. The metric option is set to 'euclidean,' indicating that the Euclidean distance will be used to calculate the distance between instances.

```
# create a KNN classifier with k=20 and use Euclidean distance as the distance metric
knn = KNeighborsClassifier(n_neighbors=20, metric='euclidean')

# train the classifier on the training data
knn.fit(X_train, y_train)
```

Figure 3.4: Creating the KNN classifier and Training the Classifier

The fit approach is used to train the KNN classifier using training data ('X_train') and labels ('y_train'). The trained classifier is used to the test data ('X_test') to create predictions. The anticipated labels are saved in the variable 'y_pred'.

Several classification metrics are computed to assess the classifier's performance. The accuracy, precision, and recall scores are calculated using the accuracy_score, precision_score, and recall_score functions, respectively. The average="macro" option is configured to compute the metrics for each class before calculating the average.

```
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average = 'macro')
recall = recall_score(y_test, y_pred, average = 'macro')
```

Figure 3.5: Evaluating KNN performance

The outcomes are displayed on the console. The accuracy, precision, and recall are all presented separately. Furthermore, the classification_report function is used to provide a complete report that includes metrics like precision, recall, F1-score, and support for each class in the test data.

```
Accuracy: 0.9098712446351931
Precision: 0.8587209302325581
Recall: 0.6910714285714286
```

Figure 3.6: Printing the results

Chapter 4

Matrices

4.1 Random Forest Classifier

4.1.1 Confusion Matrix

The confusion matrix measures the performance of a classification model by comparing predicted labels to actual labels of a collection of data. Each row in the matrix represents the actual labels, whereas each column represents the anticipated labels. The model correctly predicted class 0 for 9 samples and class 7 for 1 sample, whereas the actual label was class 0. The model properly identified 13 samples as class 1 and incorrectly predicted none. All 14 samples were accurately classified as class 2 by the model. All 54 samples were correctly classified as class 3 by the model. The model predicted incorrectly for class 4: 1 sample was classified as class 1 and 1 sample was classified as class 7, with no correct predictions for class 4. All seven samples were accurately classified as class 5 by the model. The model correctly predicted class 6 for 17 samples and class 3 for one sample, whereas the actual label was class 6. All 75 samples were correctly classified as class 7 by the model. The model accurately predicted 33 samples to be class 8 and incorrectly projected three samples to be class 9. All 32 samples were accurately classified as class 9 by the model. In summary, the confusion matrix displays the model's performance for each class, emphasizing right and wrong predictions.

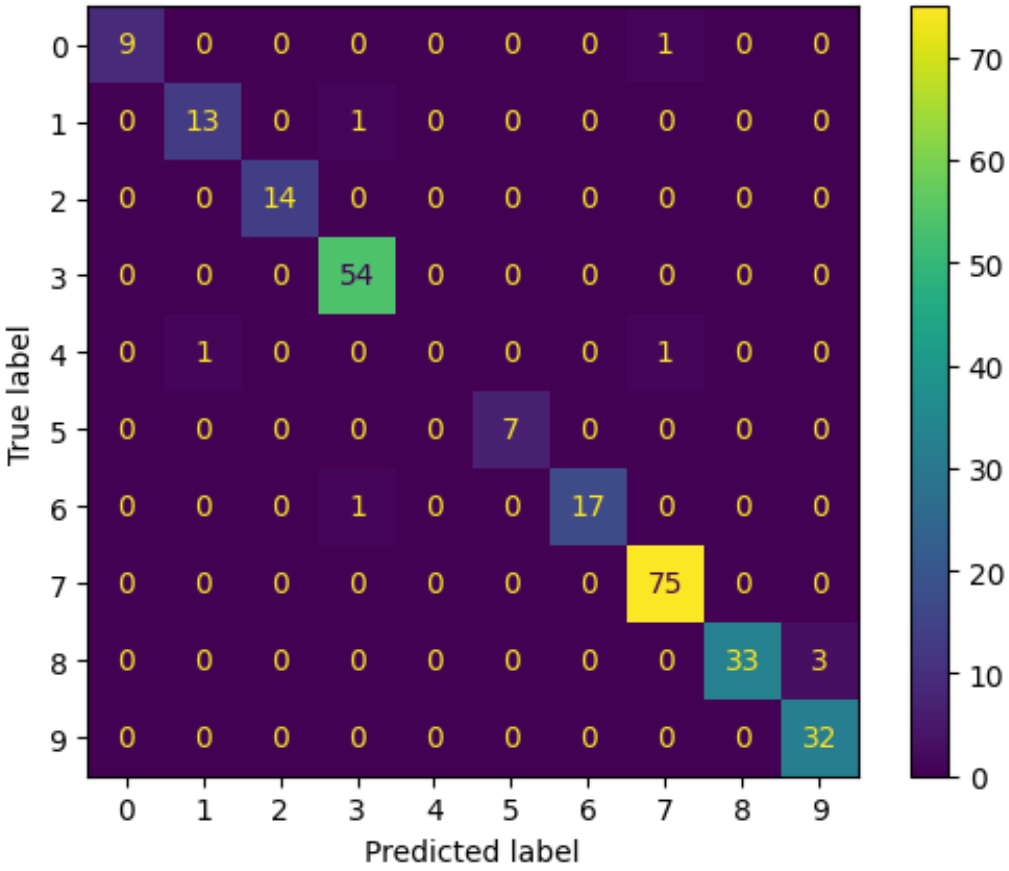


Figure 4.1: Confusion Matrix for Random Forest

4.2 K-Nearest Neighbor

4.2.1 Confusion Matrix

It is a 10x10 matrix, with each row representing the real class of the data points and each column representing the predicted class predicted by the model.

The element in the first row and first column (top left) is 0, indicating that the model correctly predicted 0 occurrences of class 0. The element in the second row and column is 3, indicating that the model accurately predicted three instances of class 1. The element in the third row and column is 12, showing that the model accurately predicted 12 instances of class 2. The element in the fourth row and column is 48, showing that the model accurately predicted 48 instances of class 3. The element in the fifth row and column is 0, suggesting that the model accurately predicted 0 instances of class 4. The element in the sixth row and column is 4, showing that the model accurately predicted four instances of class 5. The element in the seventh row and

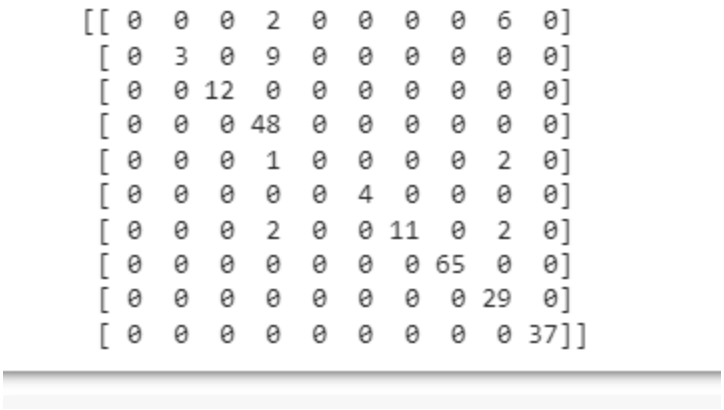


Figure 4.2: Confusion matrix for KNN

column is 11, showing that the model accurately predicted 11 instances of class 6. The element in the eighth row and column is 65, showing that the model accurately predicted 65 instances of class 7. The element in the ninth row and column is 29, showing that the model accurately predicted 29 instances of class 8. The element in the tenth row and column is 37, showing that the model accurately predicted 37 instances of class 9.

Chapter 5

Results

5.1 Results Analysis

In both random Forest and K-nearest Neghbor models, we tried with different parameter values. Below, we can show a few of the parameters we used during our experiment. For instance, we have used hyperparameter tuning in random forests to find the best hyperparameters.

5.1.1 Random Forest and K-nearest Neighbor results analysis

parameters phases	max _depth	test_size	random _state	n _estimators	cross _validation	accuracy
Phase 1	5	None	42	200	None	0.974
Phase 2	10	None	20	100	None	0.987
Phase 3	15	0.45	42	217	None	0.9770
Phase 4	5	0.45	42	100	20	0.98
Phase 5	None	0.45	42	100	None	0.96

Table 5.1: Random Forest Results in different experiments

parameters / phases	K value	precision	Recall	accuracy
Phase 1	2	0.8722	0.8270	0.9236
Phase 2	5	0.9338	0.8859	0.9656
Phase 3	10	0.7703	0.7817	0.9427
Phase 4	15	0.7636	0.7460	0.9236
Phase 5	20	0.7369	0.6666	0.8816

Table 5.2: K-nearest Neighbor Results in different experiments

Chapter 6

Conclusion

6.1 Conclusion

The most significant challenges to using DNA methylation to detect cancer origins include tumor heterogeneity, inter-tumor heterogeneity, tissue-specific signals, technical variability, a lack of comprehensive datasets, and the integration of multi-omics data. In this experiment, we utilized two classifiers to predict cancer, and we obtained significant results using Random Forest. In the future, we intend to forecast cancer using more than thirty malignancies to evaluate how accurate the prediction will be. However, we intend to introduce and test more models than the ones employed in these trials to determine which one has a strong status (results) and which one is weaker.

References

[Liu+19] Biao Liu et al. “DNA methylation markers for pan-cancer prediction by deep learning”. In: *Genes* 10.10 (2019), p. 778.

[Xia+20] Daniel Xia et al. “Minimalist approaches to cancer tissue-of-origin classification by DNA methylation”. In: *Modern Pathology* 33.10 (2020), pp. 1874–1888.

[ZX20] Chunlei Zheng and Rong Xu. “Predicting cancer origins with a DNA methylation-based deep neural network model”. In: *PloS one* 15.5 (2020), e0226461.

[Koe+21] Christian Koelsche et al. “Sarcoma classification by DNA methylation profiling”. In: *Nature communications* 12.1 (2021), p. 498.