ISSN: 0067-2904

Enhancing Early Cancer Detection: An Investigation of DNA Sequences and Machine Learning

Aadil Gani Ganie*, Samad Dadvandipour
Institute of information science, University of Miskolc, Miskolc, Hungary

Received: 2/1/2023 Accepted: 14/7/2023 Published: 30/8/2024

Abstract

The study aimed to address the global challenge of cancer-related fatalities by investigating the feasibility of identifying or predicting the early-stage presence of three distinct forms of cancers, colon, thyroid and urothelial carcinoma, via the analysis of raw DNA sequences. The data, sourced from the NCBI database, underwent a series of pre-processing techniques, including kmer analysis, under-sampling and count vectorization. Subsequently, machine learning algorithms, including logistic regression and multinomial Naive Bayes, were implemented on the pre-processed data with logistic regression demonstrating superior accuracy of 80.10% with calibration and 78.54% without calibration. To enhance the model's extrapolative capabilities, the logistic regression model was further calibrated utilizing the sigmoid method. The final model was deployed through the utilization of the open-source streamlit package.

Keywords: Cancer Classification, Machine Learning, Model deployment, Cancer identification on DNA reads.

1. Introduction

Cancer is a widespread and often fatal disease that continues to pose a significant threat to human health. The mortality and morbidity rates have been increasing over the years, and current treatment options remain insufficient, putting the lives of many patients suffering from this deadly carcinoma at risk. Late diagnosis is responsible for a significant portion of cancer-related deaths. According to World Health Organization data from 2018, there were approximately 9.56 million cancer-related deaths, and 18.08 million new cancer diagnoses worldwide with the majority occurring in low- and middle-income nations [1]. It is estimated that by 2025, there will be around 20 million new cases of cancer annually [2]. To address this issue, researchers have suggested the use of systematic methodologies based on global gene expression data to improve our understanding and classification of cancer [3][4][5]. This approach involves utilizing gene expression data, which can be obtained through the use of microarray technology, to simultaneously monitor and classify cancer.

Early detection of cancer can improve treatment outcomes, as demonstrated by a study which found that early detection of skin cancer increases treatment rates to 90% [6]. This research aimed to identify and classify cancer at an early stage using raw DNA sequences from NCBI for three types of cancer: colon, thyroid and urothelial carcinoma. The collected data contained numerous duplicate reads which were removed to improve model accuracy and reliability. While deep learning and machine learning methods have been developed for cancer

*Email: aadilganiganie@gmail.com

classification [7][8], they have not been widely implemented in practice due to their resource-intensive nature and high data, storage, and processing requirements. In this study, we employed logistic regression for cancer classification using a kmer approach, treating each DNA read as a sentence.

The language of life, as encoded by DNA and protein sequences, contains the instructions and purposes for the molecules present in all living organisms. The genome is often compared to a book, with genes and gene families as sentences and chapters, k-mers and peptides as words, and nucleotide bases and amino acids as the alphabet of this sequence language. In this study, our aim was to detect and classify three types of cancers, colon cancer, thyroid cancer, and urothelial carcinoma, at an early stage using raw DNA sequences obtained from the NCBI database. To extract features for our machine learning model, we employed techniques such as k-mer and CountVectorizer. The trained model was evaluated and deployed using the open-source package Streamlit. While our model performed well, there is still potential for improvement to increase the confidence level of classification. The code for model development and deployment is available on GitHub. To classify or detect cancer from the aforementioned types, users can simply enter a DNA read and click submit, and the trained model will classify the DNA read with a confidence interval.

2. Literature Review

Both machine learning and deep learning have been applied to various bioinformatics tasks [9][5][10][11]. Machine learning has been used in cancer classification. Authors in [12] used several machine learning methods such as K-nearest neighbor (KNN) and Naïve Bayes (NB) for breast cancer classification. KNN proved to be more successful with 97.51% accuracy than NB with 96.19%. In a separate study, the authors used the concept of machine learning with natural language processing (NLP) followed by word embedding for cancer classification between normal vs. cancer on the MIMIC III dataset [13]. They achieved an F1 score of 0.980 for cancer vs. non-cancer and an F1 score of 0.986 for breast cancer vs. other cancers. The authors of this study designed a model based on a cohort of 209 patients which demonstrated an overall area under the curve (AUC) of 0.893. The model's performance was further evaluated with respect to the stage of liver cancer revealing an AUC of 0.874 for early-stage Barcelona clinical liver cancer (BCLC stage 0-A) and 0.933 for advanced-stage BCLC (stage B-D) [14]. This study led to the conclusion that a person's risk of developing cancer can be quickly determined based on their DNA sequences. Different classifiers, including artificial neural networks, k-nearest neighbors, decision trees, fuzzy classifiers, Navies Bayes classifiers, random forests, and support vector machines, have been employed for the research [15]. In another study, the researchers modeled the variation in DNA copy number across the genome using a Bayesian hidden Markov model (HMM) with Gaussian Mixture (GM) Clustering [16]. Whereas the authors of a different study attempted to identify the mutation gene for cancer classification [17]. According to them, "dentification of a mutation in gene sequences is the preliminary step in the diagnosis of cancer." They used a support vector machine for binary data classification, i.e., cancer vs. non-cancer, and claimed an accuracy of 100%. In reference [18], the authors conducted experiments on four diverse categories of datasets and employed nine supervised machine learning techniques, including the k-Nearest Neighbors (k-NN) algorithm, Decision Trees (DT), Naive Bayes (NB), Support Vector Machines (SVM), Random Forest (RF) [19][20], AdaBoost (AB), and Gradient Boosting (GB). The decision tree machine learning technique outperformed all supervised models with the highest classification accuracy of 94.03%. We observed that most of the work had been done until model evaluation, and there were rare cases of model deployment with the type of dataset we had selected for our study.

3. Results and Discussions

We divided the results and discussion into several parts.

3.1 Data Gathering
3.2 Data Preprocessing and Feature extraction
3.3 Machine learning pipeline or model building
3.4 Model evaluation
3.5 Model deployment

3.1 Data Gathering

In this study, we focused on three types of cancers: colon cancer, thyroid cancer and urothelial carcinoma. The data used was obtained in Fastq format from the NCBI SRA database and was decompressed using the SRA toolkit. To ensure valid and reliable results, we used uniform parameters for the analysis of each cancer and data collection, including a library for every sample.:

Instrument: NextSeq550
Strategy: RNA-seq
Source: Transcriptome
Selection: cDNA
Layout: Single

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1}$$

Figure 1: Dataset description.

After decompressing the three Fastq files and extracting the sequences the data looked as below.

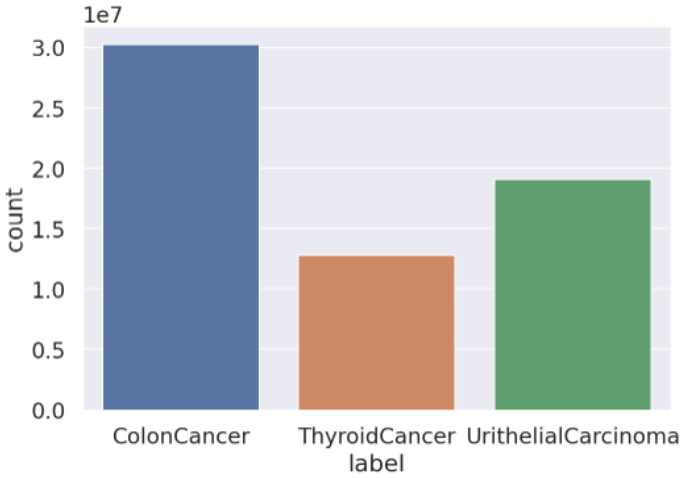

Figure 2: Decompressed data.

3.2 Data Preprocessing

The data collected was highly imbalanced, with the dominance of colon cancer sequences. To obtain reliable and generalized results for prediction or classification, the data should be as balanced as possible. There are several techniques to balance the data:

1- Oversampling: It is a technique used in data analysis and machine learning to balance the distribution of a data set by increasing the number of samples from the underrepresented class.

This is particularly useful when working with imbalanced datasets, where one class significantly outnumbers the other.

2- Undersampling: It is a technique used in data analysis and machine learning to balance the distribution of a data set by reducing the number of samples from the overrepresented class. This is useful when working with imbalanced datasets, where one class significantly outnumbers the other.

3- Downsampling: It is a technique used in data analysis and machine learning to balance the distribution of a data set by reducing the number of samples from the overrepresented class. This is useful when working with imbalanced datasets, where one class significantly outnumbers the other. For this study, down-sampling was used as oversampling would produce more duplicate reads which would affect the model performance as we could have the same samples or reads in the test data as well. The downsampling was performed on colon cancer with a factor of 0.45 and the data after the process looked more balanced than before.

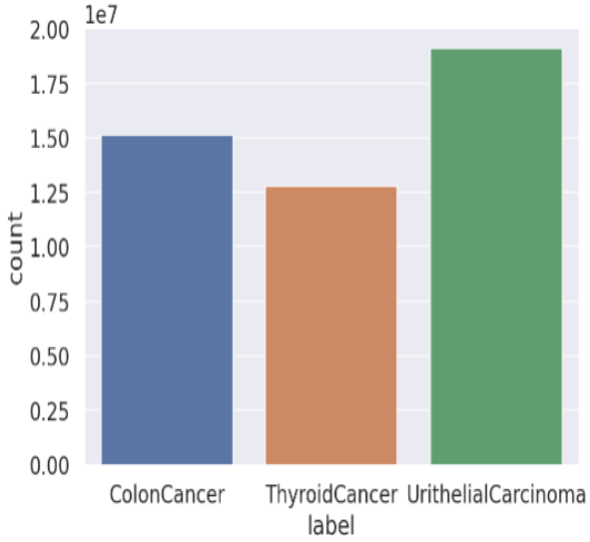

Figure 3: Number of Unique DNA sequences

All duplicate reads were removed from our data. We treated each read as a sentence so that we could apply NLP to it. K-mer was used for initial feature extraction, followed by Countvectorizer. In the field of genomics, a k-mer is a sequence of k nucleotides (A, C, G, or T) found within a DNA or RNA sequence. K-mers are used to represent the content of a DNA sequence and are often used as features in machine learning algorithms for tasks such as sequence classification and motif discovery. In natural language processing (NLP), k-mers can also be used to represent the content of a text sequence. For example, a k-mer representation of a sentence could be constructed by taking all the possible substrings of length k within the sentence. These k-mers could then be used as features for NLP tasks such as language modeling or sentiment analysis. To use k-mers in NLP for DNA sequences, we first extracted the DNA sequence from the data, then split the sequence into k-mers of length k using the string.slice() method. After extracting the k-mers from the DNA sequence, we can use them as features in a machine learning algorithm to perform tasks such as sequence classification or motif discovery.

The CountVectorizer function is used to transform a collection of text documents or DNA sequences into a document-term matrix. A document-term matrix represents each row as a specific document, and each column as a term found within the corpus. The entries within the matrix indicate the number of occurrences of each term within the corresponding document. This matrix can then be used as input for machine learning algorithms that perform tasks such as sequence classification and sentiment analysis.

3.3 Machine learning pipeline or model building

In the field of genomics, a significant amount of research has utilized deep learning techniques. However, for our classification task, a decision was made to employ a shallow machine learning algorithm, specifically logistic regression. One reason for this decision was that deep learning models can be resource-intensive, requiring substantial amounts of memory, RAM, and GPU power for training and deployment. In contrast, logistic regression is a more computationally efficient method, analogous to a "needle in a haystack" approach rather than an "axe to a tree" approach. This makes it a more suitable choice for our purposes. A sigmoid function, which is also known as the logistic function, is used in logistic regression.

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \quad (1)$$

Logistic regression can be fitted by utilizing the likelihood method due to its capability to predict probabilities. The predicted class of each training data point, represented as "y", is calculated for each input value "x". If "y" is equal to 1 or 0, the probability of "y" is represented by "p" or "1-p", respectively. The likelihood can be expressed mathematically as follows:

$$L(\alpha_0, \alpha) = \prod_{I=1}^{n} \boxdot p(x_i)^{y_i}(1-p(x_i)^{1-y_i} \quad (2)$$

By taking the log, multiplication can be converted into a summation.

$$l(\alpha_0, \alpha) \wedge \sum_{i=0}^{n} \boxdot y_i log\ p(x_i) + (1-y_i) log\ 1-p(x_i)$$
$$\sum_{i=0}^{n} \boxdot log\ 1-p(x_i) + \sum_{i=0}^{n} \boxdot y_i log\ \frac{p(x_i)}{1-p(x_i)} \quad (3)$$

To improve the reliability of our machine learning model we used *model calibration* for predictions by adjusting the model's prediction probabilities to match the observed frequency of each class in the training data.

3.4 Model Deployment

The deployment of a machine learning model refers to the process of making the model accessible and operational within production environments, such as web applications or other systems. This often involves converting the model into a format that can be easily utilized by the target environment and providing a means for accessing the model's predictions. We used Streamlit, an open-source Python library, to deploy our trained model. Streamlit simplifies the creation and deployment of interactive, web-based machine learning applications. The following images depict some examples of the performance of our model.

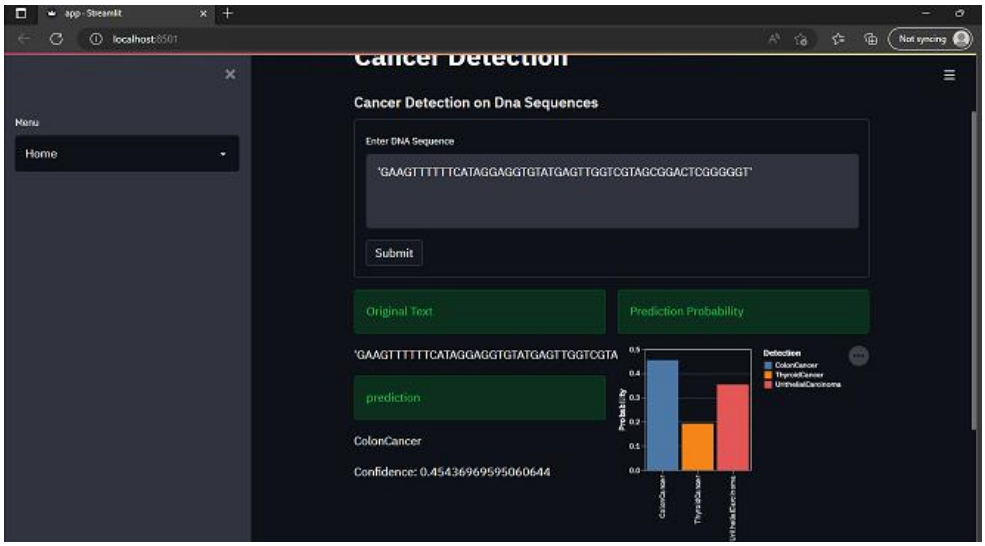

Figure 4: Colon cancer prediction on DNA sequences.

The above figure shows that our model predicted the read as colon cancer with 45% confidence. The read was taken from the colon cancer data, so the prediction was correct, however the confidence level was low.

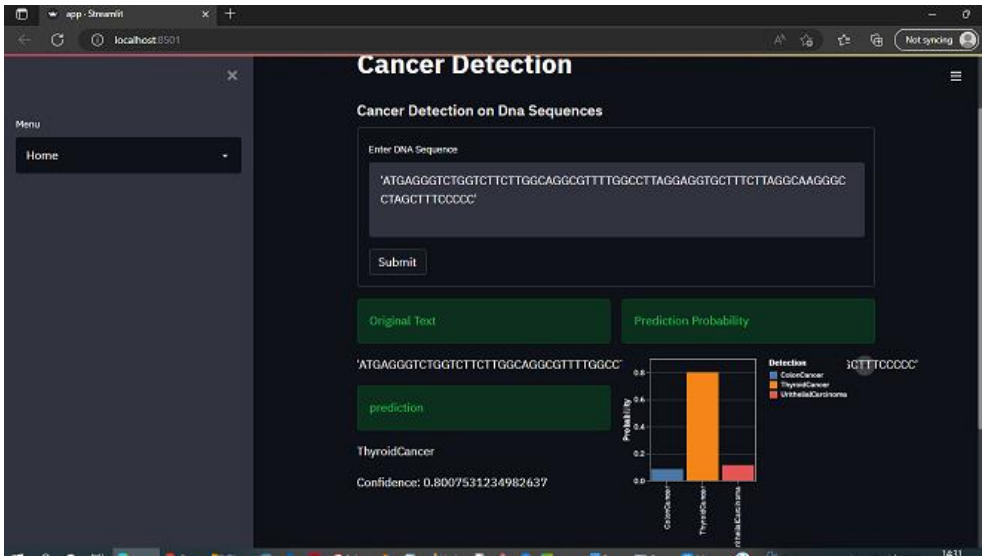

Figure 5: Thyroid cancer prediction by our deployed model.

The above read has been taken from thyroid cancer and our model predicted it as thyroid cancer with 80% confidence.

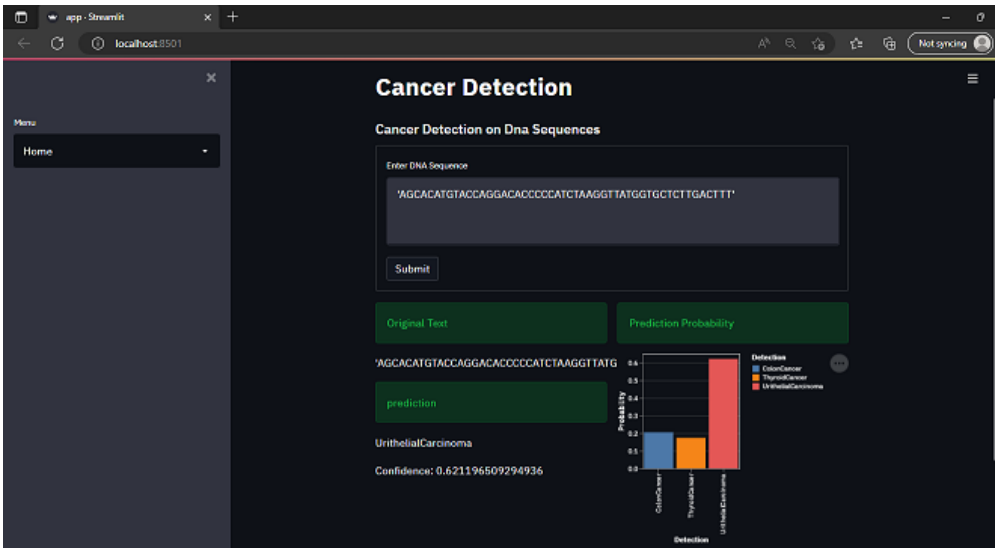

Figure 6: Urothelial carcinoma prediction on DNA sequences.

The above read belongs to the urothelial carcinoma class and our model correctly identified it with a confidence level of 62%. Although our model was generally able to predict the correct class, the confidence level was not as high as desired, indicating a need for further improvement. The primary goal of this study was to accurately classify the type of cancer using raw read sequences, and while our model showed promise, there was a potential for further enhancement. The receiver operating characteristic (ROC) curve is a popular technique for evaluating the performance of machine learning models. This curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold values. The TPR, also referred to as sensitivity or recall, quantifies the probability of correctly identifying the positive class. A model with a ROC curve closer to the upper-left corner has higher overall accuracy.

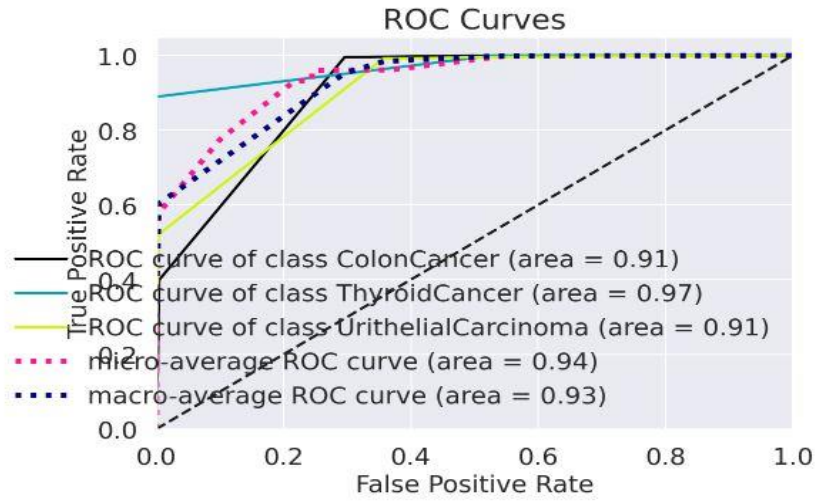

Figure 7: Logistic Regression ROC

The area under the curve for all three classes is more than 90% which reflects that our model performed well on all three types of cancers.

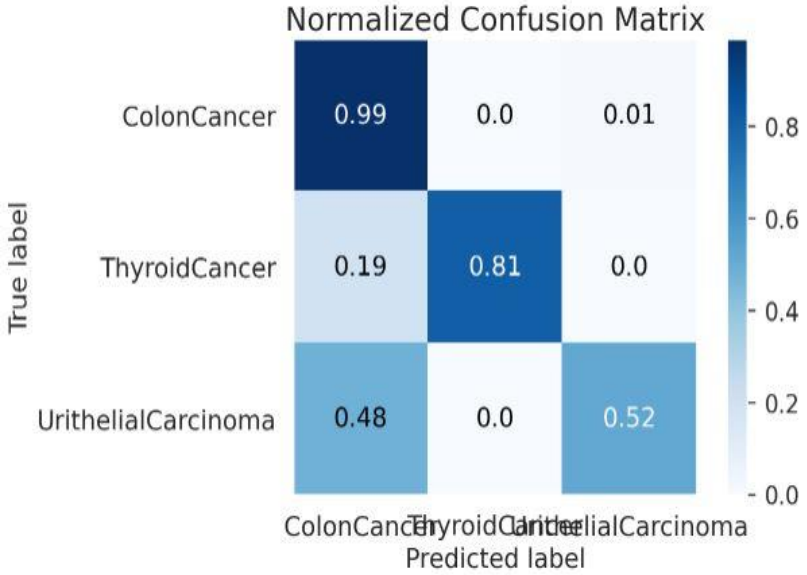

Figure 8: LR Confusion Matrix.

For colon and thyroid, the number of false positives and false negatives was very low in the Logistic Regression model. However, in the case of urothelial, it was just above 50%. Next, we applied Multinomial Naive Bayes and the ROC and confusion matrix as shown below. The micro-average was slightly lower than that of LR.

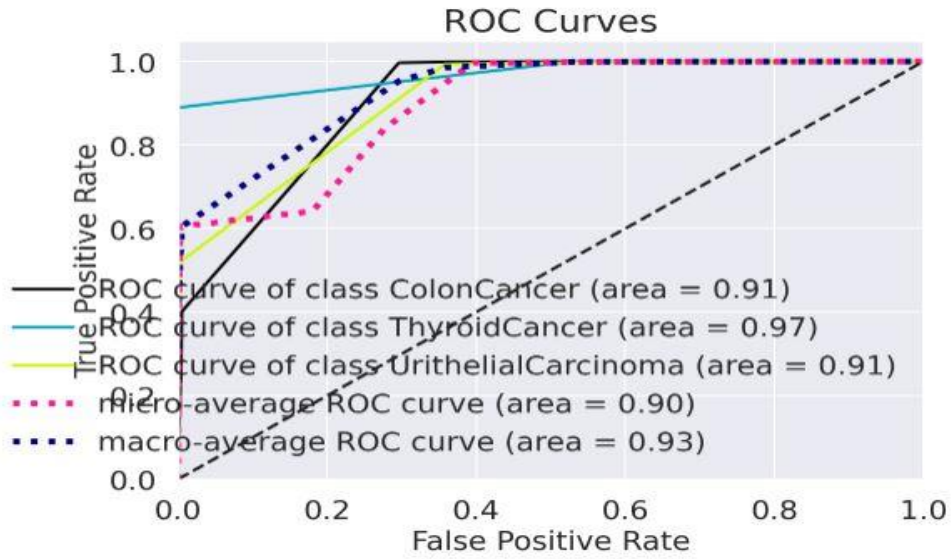

Figure 9: Multinomial NB ROC

In logistic regression, there wasn't much difference between the initial model and the calibrated model. Hence, we did not include the confusion matrix of the calibrated LR. However, in the case of multinomial Naive Bayes, there was a significant improvement in the model after calibration.

The non-calibrated multinomial NB model had below-par performance for the colon cancer class, but after calibration, its true prediction rate improved up to 99%. The accuracy for the other classes remained almost the same.

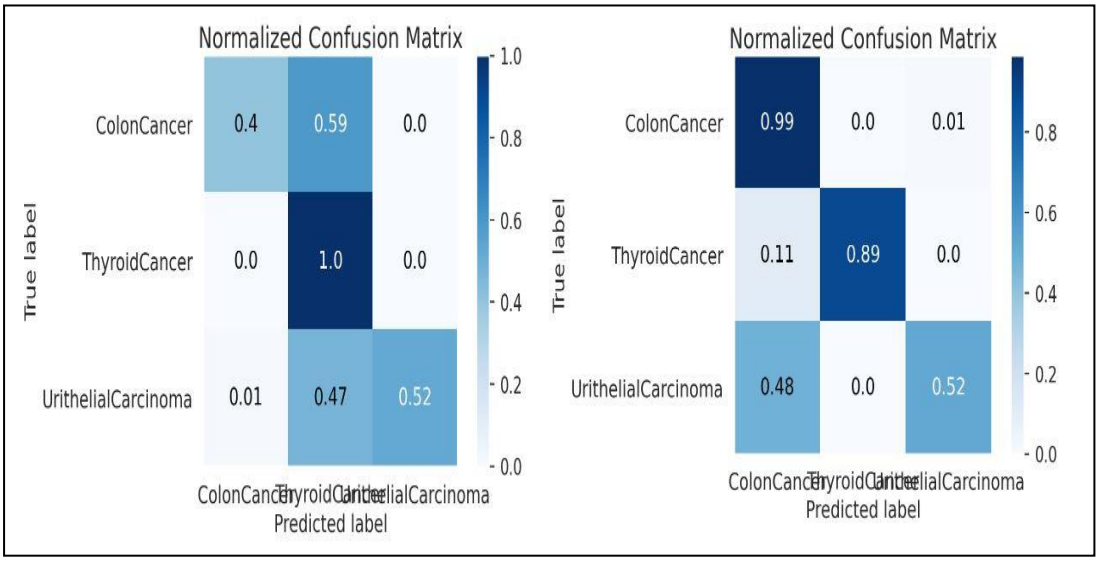

Figure 10: Multinomail NB non-calibrated vs calibrated.

4. Conclusion

The early detection of cancer has the potential to save numerous lives globally, and accurate prediction of cancer using DNA sequences can contribute significantly to this cause. In our study, we utilized machine learning algorithms to classify DNA sequences belonging to three different types of cancers. Logistic regression proved to be more effective than the multinomial NB method. In contrast to previous research, we deployed our optimal model using the Streamlit library. Model calibration is crucial for the generalization of a model, particularly for

sequential data such as DNA, where variations in the data are minimal. While our model was able to classify DNA sequences into the correct class, the confidence level was not optimal, indicating the need for further improvement to enhance the model's ability to predict DNA sequences with higher confidence. Although our study produced promising results, there is still room for improvement. The model's confidence level was not optimal, indicating the need for further refinement to enhance its ability to predict DNA sequences with higher accuracy. Future work could involve exploring alternative machine learning algorithms or utilizing additional data sources to increase the training set size and improve the model's ability to generalize to unseen data.

5. Acknowledgements

The success of this research would not have been possible without the dedication and hard work of the team of scientists involved.

6. Ethical Responsibilities of Authors

Authors complied with ethical standards and guidelines in this research.

7. Statements on Compliance with Ethical Standards and Standards of Research Involving Animals

This authorship did not encompass any examination that utilized animals in their methodology.

8. Disclosure and Conflict of Interest

The authors affirm that they have no conflict of interest.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, 2018, doi: 10.3322/caac.21492.

[2] J. Ferlay *et al.*, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *Int. J. Cancer*, 2015, doi: 10.1002/ijc.29210.

[3] Ganie, A. G., & Dadvandipour, S. (2022). Identification of online harassment using ensemble fine-tuned pre-trained Bert. Pollack Periodica.

[4] Thambi, R., Kandamuthan, S., Vilasiniamma, L., Abraham, T. R., & Balakrishnan, P. K. (2017). Histopathological analysis of brain tumours - A seven-year study from a tertiary care centre in South India. Journal of clinical and diagnostic research: JCDR, 11(6), EC05.

[5] Hwang, S. M. (2020). Classification of acute myeloid leukemia. Blood research, 55(S1), S1-S4.

[6] N. C. F. Codella *et al.*, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM J. Res. Dev.*, 2017, doi: 10.1147/JRD.2017.2708299.

[7] A. Kishk *et al.*, "A Hybrid Machine Learning Approach for the Phenotypic Classification of Metagenomic Colon Cancer Reads Based on K-mer Frequency and Biomarker Profiling," 2019, doi: 10.1109/CIBEC.2018.8641805.

[8] S. Liu *et al.*, "Finding new cancer epigenetic and genetic biomarkers from cell-free DNA by combining SALP-seq and machine learning," *Comput. Struct. Biotechnol. J.*, 2020, doi: 10.1016/j.csbj.2020.06.042.

[9] Y. Cao, T. A. Geddes, J. Y. H. Yang, and P. Yang, "Ensemble deep learning in bioinformatics," *Nature Machine Intelligence*. 2020, doi: 10.1038/s42256-020-0217-y.

[10] I. Mondal, Y. Hou, and C. Jochim, "End-to-End NLP Knowledge Graph Construction," 2021.

[11] W. K. Sari, D. P. Rini, and R. F. Malik, "Text Classification Using Long Short-Term Memory with GloVe Features," *J. Ilm. Tek. Elektro Komput. dan Inform.*, 2020, doi: 10.26555/jiteki.v5i2.15021.

[12] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," *J. Pers. Med.*, 2021, doi: 10.3390/jpm11020061.

[13] A. A. R. Magna, H. Allende-Cid, C. Taramasco, C. Becerra, and R. L. Figueroa, "Application of Machine Learning and Word Embeddings in the Classification of Cancer Diagnosis Using Patient

Anamnesis," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3000075.

[14] K. Tao *et al.*, "Machine learning-based genome-wide interrogation of somatic copy number aberrations in circulating tumour DNA for early detection of hepatocellular carcinoma," *EBioMedicine*, 2020, doi: 10.1016/j.ebiom.2020.102811.

[15] F. Hussain, U. Saeed, G. Muhammad, N. Islam, and G. S. Sheikh, "Classifying Cancer Patients Based on DNA Sequences Using Machine Learning," *J. Med. Imaging Heal. Informatics*, 2019, doi: 10.1166/jmihi.2019.2602.

[16] G. Manogaran, V. Vijayakumar, R. Varatharajan, P. Malarvizhi Kumar, R. Sundarasekar, and C. H. Hsu, "Machine Learning Based Big Data Processing Framework for Cancer Diagnosis Using Hidden Markov Model and GM Clustering," *Wirel. Pers. Commun.*, 2018, doi: 10.1007/s11277-017-5044-z.

[17] D. W. Liu *et al.*, "Automated detection of cancerous genomic sequences using genomic signal processing and machine learning," *Futur. Gener. Comput. Syst.*, 2019, doi: 10.1016/j.future.2018.12.041.

[18] B. Kurian and V. L. Jyothi, "Breast cancer prediction using an optimal machine learning technique for next generation sequences," *Concurr. Eng. Res. Appl.*, 2021, doi: 10.1177/1063293X21991808.

[19] Al-Jaburi, A. A., & Al-Sudani, A. H. (2023). Hybrid Techniques with Support Vector Machine for Improving Artifact Ultrasound Images. *Iraqi Journal of Science*, 944-957.

[20]Nasser, F. K., & Behadili, S. F. (2022). Breast Cancer Detection using Decision Tree and K-Nearest Neighbour Classifiers. *Iraqi Journal of Science*, 4987-5003.