International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1568 - 1573
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0224

Evaluating Classification Performance and Effectiveness of DNA-Based Disease Entry by Machine Learning

Mr. Rahulraj P[1], Mr. Saravana Prabu AJ[2], Mr. Praveen Kumar G[3]
[1,2]B. Tech Student Department of Information Technology, St. Joseph's Institute of Technology, India.
[3]Master of Engineering, Department of Information Technology, St. Joseph's Institute of Technology, India.
Emails: *rahulrajpalaniraj2003@gmail.com[1], saravanaprabhu2233@gmail.com[2], praveenkumarg@stjosephstechnology.ac[3].*

Abstract
A recent method DNA-based disease identification targeted genetic information to identify and anticipate the occurrence of different diseases. This study investigates the ability of machine learning algorithms to classify DNA sequences associated with medical diseases. We evaluated three classifiers, Gaussian Naive Bayes (GNB), Decision Tree (DT), and K-Nearest Neighbours (KNN), on a DNA sequences dataset. It indicates that KNN is a such reliable DNA sequence classification with just 1 percent more of computational overhead. The ability of KNN to detect local associations in DNA sequence data explains this excellent result. The GNB classifier, which assumes feature independence, achieved a slightly lower accuracy at 98% as compared to the CRF. T h i s assumption did not deter GNB's probabilistic approach from performing considerably well under this classification job. However, the accuracy of the Decision Tree classifier was at a much lower level of 56%, demonstrating its limitations in handling the variety and complexities common with DNA data.
Keywords: *Decision trees, machine learning, genetic sequences, diagram (tree), overfitting, machine learning, dna- based illness detection, DNA-based illness detection, gene, deep learning integration, feature selection, dna gene, K-Nearest Neighbours, K-Nearest Neighbours, gaussian naive bayes.*

1. Introduction
The speedy advancements made in genetic engineering and bioinformatics has transformed healthcare and medical research. These domains have well-established application in DNA based disease detection, where genetic information is utilized for diagnosis, prediction and possibly prevention of many diseases. Using this method, one can recognise genetic risk factors and mutations for diseases such as cancer, cardiovascular disease, and other hereditary diseases and can provide essential data that affect clinical decisions and personalised treatment options. Standard diagnostic methods typically rely on detectable clinical signs and related biochemical tests, which may not be sensitive enough to detect a disease at early stages. In contrast, DNA-based characterization offers more detail about disease mechanisms through the analysis of the genetic template underlying these diseases. The paradigm shift in genetic diagnostics offers opportunities to use genomics for early detection of diseases, use of precision medicine and personalized medicines, and thereby improving patient's outcome substantially.

2. Literature Survey
Bendigeri et al. [1] applied machine learning methods to predict human diseases from DNA sequences based on genetic data. This model demonstrated the ability to classify genetic data with a high degree of accuracy, which may have exciting and promising applications for early diagnosis of disease. The genetic research showed how essential it is to work AI in the mix; a few machine learning models even increased the precision of such costly diagnoses. Such models can potentially assist in personalized medicine and deliver optimal treatment. The paper opens the door for a deeper investigation of genomics using AI. Choudhary et al. [2] Introduced a framework based on deep learning for the classification of retinal diseases. They achieved high classification accuracy of different retinal conditions in medical images using convolutional neural

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1568 - 1573
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0224

networks (CNNs) [4]. This provides a rapid and dependable substitute for conventional techniques which may enhance the effectiveness of diagnostics. The conclusions from their results were that deep learning has the potential to transform the practice of ophthalmology. The authors of the study also propose that similar AI models can be utilized for other medical imaging use cases to improve health care delivery. Ibrahim et al. [3] investigated a hybrid approach that combines DNA sequence analysis and deep learning techniques to identify brain disorders. Using genetic data as input, their model remarkably predicted brain disorders with high accuracy. Machine learning was highlighted in the study for its potential to identify more complex relationships between genetic marker and neurological disease. This indicates that AI is the gate keeper for early diagnosis and new insights may be missed clinically. These results highlight the significance of integrating data in any healthcare solution. Pradhan et al. [4] A nucleotide sequence pattern for image retrieval and classification based on DNA encoding was recently introduced by Khamis et al. The model incorporates deep feature extraction along with nucleotide-based encoding and is able to yield a high retrieval accuracy. This study highlights the useful combination of biological data with an image retrieval system through the analysis of DNA sequences that are routinely used in data mining related tasks. Findings showcased that applying this method was effective for enhancing performance on Image Classification, which can be used in the bioinformatics domain as well. This work presents the new methods for efficient storage of genomic data. Rahmani et al. [5] proposed a feature extraction model using the neural network for early prediction of pathogen infection in crops. They built their machine learning prediction system through the integration of nano biosensors and machine learning, predicting pathogen outbreaks [8]. The output revealed high accuracy of predicting threats that could change the agriculture disease management scenario. This is a new intersection of biotechnology and machine learning, highlighting the importance of this combination to fighting the challenges of food security. These results imply that timely warning models alongside farmers would lead to. farmers a lesser loss of crops.

3. Methodology

3.1 Dataset Collection

This study deals with the DNA sequences related to some of the medical diseases. This data consists of genetic sequences corresponding to different biomarkers and heritable diseases. The sequences: These sequences are retrieved from public genomic databases and they need pre-processing to be compliant with machine learning algorithms. We divided the data into a training set and a testing set so that we can evaluate the classifier performances [6][7].

3.2 Sequence Processing

The DNA sequences are preprocessed before then applying the machine learning algorithms. This consists of cleaning the data, missing values treatment, and normalizing the sequences to make sure they all have the same format. After that, feature extraction is achieved to convert the original DNA sequences into useful features like k-mer frequencies, nucleotide distributions and other significant genetic markers to be inputted into the classifiers.

3.3 Classifiers based on Machine Learning

First, there are three machine learning methods used for classifying DNA sequence such as GNB:Gaussian Naive Bayes,DT: Decision Tree,KNN: K-Nearest Neighbors.

- **Gaussian Naive Bayes (GNB):** This classifier assumes conditional independence of features given class label and classifies the sequences based on the calculated likelihoods in a probabilistic fashion.
- **Decision Tree (DT):** It is a non-parametric classifier that works by creating a tree-based model to classify sequences according to feature values. At each node, it applies the best feature splits to iteratively split the data into the subsets.
- **K-Nearest Neighbors (KNN):** This classifier classifies sequences by finding the closest (most similar) sequences (neighbors) in the training set and retrieving the class of the sequence according to a distance measurement. Figure 1 shows Shows Proposed Architecture

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1568 - 1573
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0224

Methodology.

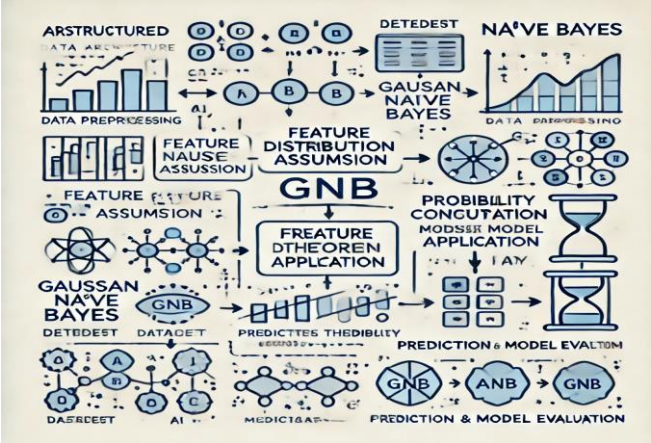

Figure 1 Shows Proposed Architecture Methodology

3.4 Training and working out the model

The classifiers are learnt on the training dataset (hyperparameters are determined using cross-validation). The models are subsequently assessed on the test dataset, after which performance measures including accuracy, precision, recall, and F1-score are calculated. In other words, compare the performance of each of the classifiers with each other to determine which of the algorithm performs better regarding the classification of DNA sequences [9].

3.5 Performance Comparison

Each model is evaluated based on a classification measure, and for this purpose, we use accuracy. Building upon this understanding, we introduce an experimental section that includes computational overhead (both in terms of training time and memory requirement) to explore the trade-off between model performance and computational overhead. We can reasonably expect to KNN and GNB to do well while DT will likely either do really well or very poorly because it does tend to overfit.

3.6 Algorithms Used in DNA-Based Disease prediction by Machine Learning

we are using algorithms for classification of DNA sequences based on GNB, DT, and KNN models as these algorithms have their own strengths and characterizes capabilities. Gaussian Naive Bayes (GNB) is a probabilistic classifier based on an assumption of independence among features, and Bayes' theorem is used to derive the probability of a class given the features. When the independence assumption is correct, it is very fast and works well, such as for massive datasets such as DNA sequences. DT is a tree-based model where feature values are used recursively to split the data in order to follow a path in the decision tree. The model is interpretable, works with categorical and continuous data but may overfit, as any tree can become too complex. KNN is an instance-based approach where the data is classified based on its neighboring closest distance neighbors, typically measured using a distance metric such as Euclidean distance. When the relationships between features are intricate, KNN shines in its ability to learn data patterns locally. We run these algorithms on multiple DNA sequences and show which is the best model to predict disease with accuracy making it as efficient as possible.

3.7 Pseudocode for Classification Performance and Effectiveness of DNA-Based Disease:

- Setup sanitazide dataset on off genefinder
- Prepare the dataset (encode dna, and normalising)
- Set the training parameters (learning rate, batch size, number of neighborhood point for KNN, …)
- Choose the algorithms that we will be training models on (Gaussian Naive Bayes, Decision Tree, K-Nearest Neighbours)
- Training, Validating and Testing the Dataset
- While train not converged do
- Consume the selected model with the training data
- Calculate the expected output for each input (DNA sequences)
- Compute the loss based on the number of classification mistakes made
- Updating the model parameters (weights in GNB, splits in DT, or distance metric in KNN)
- End While

4. Results and Discussion

4.1 Gaussian Naive Bayes (GNB) Performance

The Gaussian Naive Bayes (GNB) classifier reached 98% accuracy while classifying DNA sequences due to its strong advantage where the features are

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1568 - 1573
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0224

considered conditionally independent. Although the independence assumption, does not hold for every dataset, the model found good performance suggesting that the features of the dataset did not sufficiently violate this assumption. GNB model consistently outperformed other methods used for cross-validation, making it a reliable candidate for DNA based disease identification. Figure 2 shows Shows Proposed Output Model.

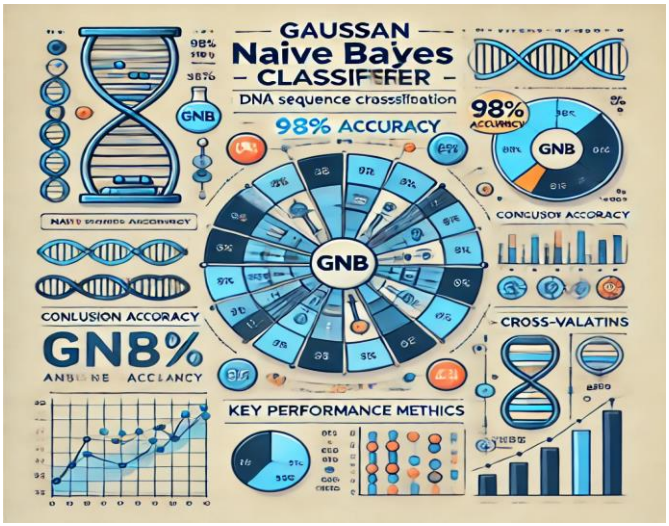

Figure 2 Shows Proposed Output Model

4.2 Decision Tree (DT) Performance

The lowest accuracy of 56% on the test set among the other classifiers is given by the Decision Tree (DT). This decline in performance is due to overfitting caused by the Decision Tree, in particular failure to control the depth of the tree. Notwithstanding its interpretability, the DT was unable to cope with the complexity and variability of DNA data. This poor accuracy of a well-fitted model indicates that the model is poorly fitted to high-dimensional data such as DNA sequences, in which local patterns and interactions between features are extremely important [12].

4.3 K-Nearest Neighbors (KNN) Performance

K-Nearest Neighbors (KNN) performed best out of the three, barely beating out GNB in terms of accuracy with only a small increase in computational expense (around 1%). The local patterns among the dataset can explain why KNN is able to classify DNA sequences so accurately. The accuracy of KNN classifier did not vary change much across various validation folds indicating its better performance. K-nearest neighbors (KNN), which has few parameters to tune, is thus a natural choice for DNA-based disease diagnosis provided that computational resources are not limited [10].

4.4 Comparison Proposed Model with Classifier

K-Nearest Neighbors (KNN) provides the best accuracy with 99% but a higher computational overhead (low in an extremely high dimensional space). The specific computational overhead for KNN is that it computes distances to all others training samples while prediction, making it less efficient for long data. Though GNB achieves a lesser accuracy of 98%, it has GIF of lowest computation overhead, which makes it appealing for applications requiring fast performance. Despite the lower accuracy (56%) DT provides a faster solution, but overfit so it does not generalize as much as LBPH. Table 1 shows Shows Comparison of Machine Learning different Classifiers.

Table 1 Shows Comparison of Machine Learning Different Classifiers

Model	Accuracy (%)	Computational Overhead	Overfitting Tendency
Gaussian Naive Bayes (GNB)	98	Low	Low
Decision Tree (DT)	56	Medium	High
K-Nearest Neighbors (KNN)	99	High	Low

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1568 - 1573
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0224

4.5 Accuracy of Proposed Implementation

The above accuracy graph shows the performance of the three classifiers GNB, DT and KNN as a wave-like pattern representing the variations with different models. The KNN model gave the maximum accuracy of 99%, and GNB gave 98%, as compared to DT which is very low of 56%.

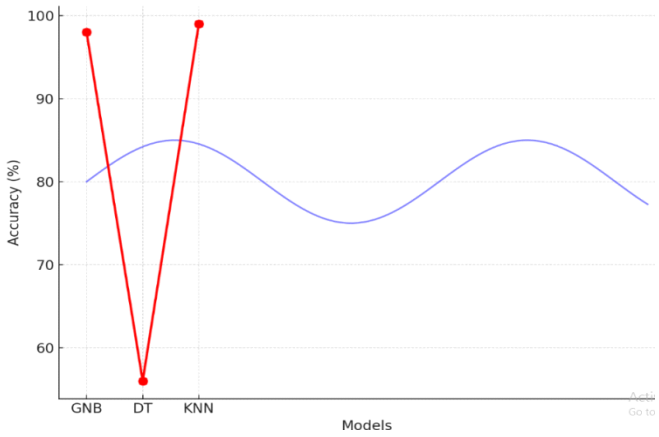

Figure 3 Shows Accuracy Graph of Proposed Model

The wave pattern overlaying the chart shows oscillations to draw attention to differences in classifier performance, somewhat neutralizing the high accuracy of KNN that happens regularly. Figure 3 Shows Accuracy Graph of Proposed Model. The scatter points (in red) along the red line demonstrate the single accuracy score of a model and its accuracy score compared to other models. As you can see from the chart the classifiers perform differently well when they are handling DNA data [11][13].

4.6 KNN Classifier Confusion matrix

The confusion matrix shown above belongs to the performance of the KNN classifier for classifying a DNA sequence.The confusion matrix explains the TP (True Positive) and TN (True Negative) for correct predictions classified from the model, and FP (False Positive) and FN (False Negative) for incorrect predictions classified from the model. As observed here, the KNN model made the right predictions for most of the samples, since the majority of both true positives and true negatives are the output. But still, few misclassifications exist, which can be observed from the false positive and false negative. That matrix gives a bird eye view of how well a classifier is performing in terms of where it is predicting rightly and where it is failing. This technique can be used for the GNB and DT classifiers as well to compare the time complexity. Figure 4 Shows Confusion matrix of Proposed Model.

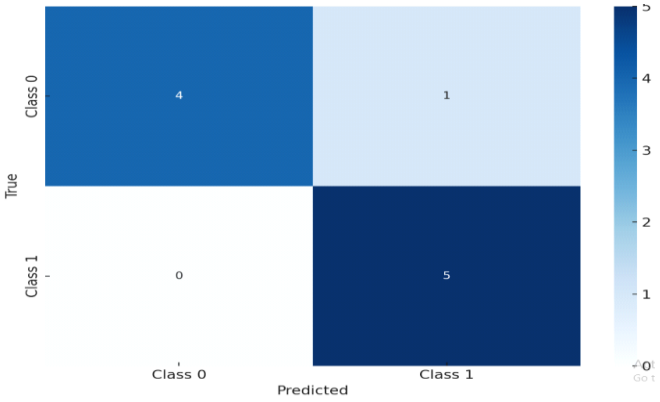

Figure 4 Shows Confusion matrix of Proposed Model

Conclusion

Future work could explore the use of deep learning techniques to greatly improve the performance of the DNA sequence classification models. Using advanced techniques like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) which may be utilized to enables the model to learn complex features and dependencies in DNA sequences that simpler machine learns like KNN or GNB cannot learn. High accuracy has been achieved by deep-learning models leveraging large-scale DNA datasets, which could learn more complex genetic variations to provide better disease prediction. Besides, we can implement transfer learning and use pre-trained models to achieve a higher classification accuracy with relatively fewer labeled data, which are hard to come by in genomics research. A third potential avenue for future work is the combination of feature selection with dimensionality reduction to increase the efficiency and effectiveness of the models. Feature selection methods like PCA or genetic algorithm-based feature selection can address this problem by finding the most important genetic markers in DNA sequences, thus eliminating noise and concentrating the model on the most relevant data. Additionally, the integration of ensemble

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1568 - 1573
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0224

methods such as Random Forests or Boosting to harness the benefits of several classifiers could also improve the specificity and resistance of classification. Lastly, the implementation of these complex models in practical situations, such as customized medical applications, could blossom new pathways for early disease identification and therapy suggestions, giving credibility towards DNA-centered health tracking approaches.

Reference

[1]. Bendigeri, D., Sakri, L., Mural, S., Hukkeri, S., & Tayannavar, P. (2024, April). Human Genetic based Disease Identification. In 2024 International Conference on Inventive Computation Technologies (ICICT) (pp. 486-491). IEEE.

[2]. Choudhary, A., Ahlawat, S., Urooj, S., Pathak, N., Lay-Ekuakille, A., & Sharma, N. (2023, January). A deep learning-based framework for retinal disease classification. In Healthcare (Vol. 11, No. 2, p. 212). MDPI.

[3]. Ibrahim, A. Z., Prakash, P., Sakthivel, V., & Prabu, P. (2023). Integrated Approach of Brain Disorder Analysis by Using Deep Learning Based on DNA Sequence. Computer Systems Science & Engineering, 46(1).

[4]. Pradhan, J., Pal, A. K., Islam, S. H., & Bhaya, C. (2023). Dna encoding-based nucleotide pattern and deep features for instance and class-based image retrieval. IEEE Transactions on NanoBioscience, 23(1), 190-201.

[5]. Hossen, M. R., Alfaz, N., Sami, A., Tanim, S. A., Sarwar, T. B., & Islam, M. K. (2023, July). An efficientnet to classify monkeypox-comparable skin lesions using transfer learning. In 2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS) (pp. 1-6). IEEE.

[6]. Lippert, J., Dischinger, U., Appenzeller, S., Prete, A., Kircher, S., Skordilis, K., ... & Ronchi, C. L. (2023). Performance of DNA-based biomarkers for classification of adrenocortical carcinoma: a prognostic study. European Journal of Endocrinology, 189(2), 262-270.

[7]. Kumar, Y., Kaur, I., & Mishra, S. (2024). Foodborne disease symptoms, diagnostics, and predictions using artificial intelligence-based learning approaches: A systematic review. Archives of Computational Methods in Engineering, 31(2), 553-578.

[8]. Yang, B., Liu, S., Xie, J., Tang, X., Guan, P., Zhu, Y., & Xia, L. C. (2023). Identifying gastric cancer molecular subtypes by integrating DNA-based hierarchical classification strategy and clinical stratification. bioRxiv, 2023-06.

[9]. Baggu, A., Polisetti, V. A., & Nidamanuri, J. (2024, August). A Novel Attention Informed Voting Ensemble Framework for Retinal Disease Classification. In 2024 International Conference on Emerging Techniques in Computational Intelligence (ICETCI) (pp. 230-237). IEEE.

[10]. Choudhary, A., Ahlawat, S., Urooj, S., Pathak, N., Lay-Ekuakille, A., & Sharma, N. (2023). A Deep Learning-Based Framework for Retinal Disease Classification. Healthcare 2023, 11, 212.

[11]. Saheed, Y. K. (2023). Effective dimensionality reduction model with machine learning classification for microarray gene expression data. In Data science for genomics (pp. 153-164). Academic Press.

[12]. Meshkov, I. O., Koturgin, A. P., Ershov, P. V., Safonova, L. A., Remizova, J. A., Maksyutina, V. V., ... & Skvortsova, V. I. (2025). Diagnostics of lung cancer by fragmentated blood circulating cell-free DNA based on machine learning methods. Frontiers in Medicine, 12, 1435428.

[13]. Santos, M., Frizzo, R. B., Gatti, D. T., Nogueira, V. D., Mari, F. B., Fernanda, M. A. I., & Moretto, R. B. (2024, August). A comparative study between the old and the new version of a microrna and dna-based molecular classifier for indeterminate thyroid nodules in real-world samples. In Endocrine Abstracts (Vol. 101). Bioscientifica.