*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 10 Issue 6 June 2022

Development of a Robust DNA Sequencing Multiclass Hybrid Classifier Employing Machine Learning and NLP For Automated Gene Classification

Vanshika Gupta [(1)], Reetu Jain[(2)]

[(1)]Student, Class 11th, ITL Public School, Sector 9, Dwarka, New Delhi-110077.
Email-id: itlschool@yahoo.in

[(2)]Mentor, On My Own Technology Pvt. Ltd., 1018, Samartha Aishwarya, Oshiwara, Mumbai, India.
Email-id: reetu.jain@onmyowntechnology.com

Abstract: Sequencing DNA approaches figuring out the order of the quad chemical constructing blocks - called bases - that make up the DNA molecule. The collection makes scientists take notice of the sort of genetic statistics which are carried in a specific DNA segment. Scientists can use collection statistics to decide which stretches the DNA to incorporate genes and which stretch deliver regulatory instructions, turning genes on or off. In addition, and more importantly, sequencing the data can spotlight modifications in a gene which could cause disease. In this study, we have built a hybrid ensemble machine learning model using eleven classifiers for the development of a DNA sequence genomic classifier using machine learning and genomic sequencing in NLP through k-mers for the best possible results. We constructed a hybrid model using NLP and machine learning to progress the gene classification rapidly, which helped in increased research of biomedical science. In the later part of the study, we built a hybrid model employing stack ensemble methodology. To assess the model's performance, we employed several performance evaluation measures. The hybrid model built on eleven classifiers gave an extremely high accuracy for three different sets of data, i.e., human, chimpanzee and dog. We did a comprehensive study on all three species, and investigated their genome classifications, to build a best fit model for them. The model for human data, reached 98.08% accuracy and an almost perfect AUC score. 90.75% accuracy was received for chimpanzee data, while 70% for dog data, making our hybrid ensemble transgress species.

Keywords: Machine learning, DNA sequencing, NLP, Gene, Classification, Model, Ensemble

1. Introduction

The gene is the fundamental unit of inheritance in almost all organisms present in this world, and the ultimate determinant of all phenotypic qualities expressed in a living being. The deoxyribonucleic acid (DNA) of an average human cell contains roughly 30,000 to 120,000 genes, however, only a minute amount of them is expressed in the gene [1]. In research, genomic DNA is a very useful tool in applications of Polymerase Chain Reaction (PCR), library construction, southern blotting, hybridizing etc. Every quality comprises a few sections with different purposes of every one of them, each associated with an alternate piece of the course of articulation. The fundamental rehashing units of the DNA polymer is nucleotides. Nucleotides comprise of an invariant piece, a five-carbon deoxyribose sugar with a phosphate bunch, and a variable part, the base. Of the four bases that show up in the nucleotides of DNA, two are purines, adenine (A) and guanine (G), and two are pyrimidines, cytosine

(C) and thymine (T). Nucleotides are associated with one another in the polymer through their phosphate gatherings, leaving the bases allowed to cooperate with one another through hydrogen holding [2].
Natural language processing (NLP) is an algorithm-based guideline that makes a difference within the examination of content information, pictures, and other unstructured data. It summarizes the total dataset from which one can at that point create significant bits of knowledge. Nowadays, most of the data in healthcare is unstructured. This incorporates data within the shape of journals, enlightening, blogs, and social media intuitively. In a situation like healthcare, it makes more sense to tap into the potential of unstructured information. [3]

Machine learning is a domain of artificial intelligence, which in simple terms, said to be the capacity or ability of a machine to be able to copy human intelligence behaviors, machine learning enables computers to be able to perform tasks in the same way as humans do.

Several researchers in the past have employed machine learning and NLP in genomic science. F. Khan et al., in 2020, proposed a DNA Act-Ran - a Digital DNA Sequencing Engine for Ransomware Detection, using machine learning, to understand the techniques used in ransomware development and their deployment for better counter measures [4]. They achieved an accuracy ranging from 75.81 to 87.91 percent [4].

Adam L Bazinet and Michael P Cummings in 2012 proposed a comparative evaluation of sequence classification programs by dividing the huge pre-existing programs into defined categories based on the algorithm they made use of [5]. They were unable to analyze most of the datasets with fluorescence-activated single cell sorting (FACS) and found out (Metagenome Analysis Software) MEGAN gave the most accuracy while MetaPhyler had the least running time [5].

K.Vervier et al. in 2016 proposed a large-scale machine learning for metagenomics sequence classification by using a machine learning based approach for taxonomic assignments of metagenomic reads [6]. They received a precision of up to 97.5% [6].

Moreover, H. Gunasekaran et al. in 2021 examined DNA classification for mathematical aspects behind deep learning and transfer learning approaches for medical image analysis using (convolutional neural network) CNN and hybrid models. From their experimental results, they received an accuracy of 93.16% and 93.13%, respectively, on testing data [7].

Therefore, in this study, we aimed to analyze the pattern of the nucleotide by developing a robust DNA sequencing multiclass hybrid classifier employing NLP and machine learning for automated gene classification.

To build the model, we used genome data of three different species: human, dog, and chimpanzee and proposed to build a classifier whose main aim is to predict a gene's function based on DNA sequence. There were seven classes of gene which the model was trained upon including: G protein coupled receptors, Tyrosine kinase, Tyrosine phosphate, Synthetase, Synthase, ION channel and transcription factor. We used twelve machine learning classifiers i.e., logistic regression, decision tree, random forest, kNN, multinomial naive bayes, SVM, linear SVC, passive aggressive, Bernoulli NB, extra tree, bagging and gradient boosting as a base classifier. In the later part of the study, we built a hybrid model employing stack ensemble methodology. To assess the model performance, we employed several performance evaluation measures.

Thus, by constructing a hybrid model ensemble (stacking with eleven different classifiers) using NLP and machine learning helped us progress the classification of the seven genome types rapidly. This would easily help further the field in biomedical research, and to equivocally find out more about genes.

2. Methods

2.1 Workflow Design

2.1.1 NLP in Genetics

NLP is a domain of artificial intelligence which is mainly focused with the interactions between computers and human language, in particular, how to program computers to process and analyze large amounts of human language and convert into a form that is understandable and readable by them. Biological NLP is a field of research that aims to systematically investigate the dynamics between genes, and sequences, forming the foundation for machine learning biomedical literature. [8]

2.1.2 Machine Learning in Genetics

Machine learning is extremely useful in identifying the genetic factors hidden beneath for diseases and gene classifications by searching for genetic patterns amongst people by comparing and re-assessing their genetic patterns. Many new advancements in the field of human genetics can be credited to machine learning.

Thus, this study aims at achieving highly accurate predictions for models to be able to depict patterns through DNA sequences using hybridization of models.

Figure 1 presents the workflow diagram of the performed study. It showcases the process of the construction of the automated hybrid model for gene classification and its employment in the classification system.

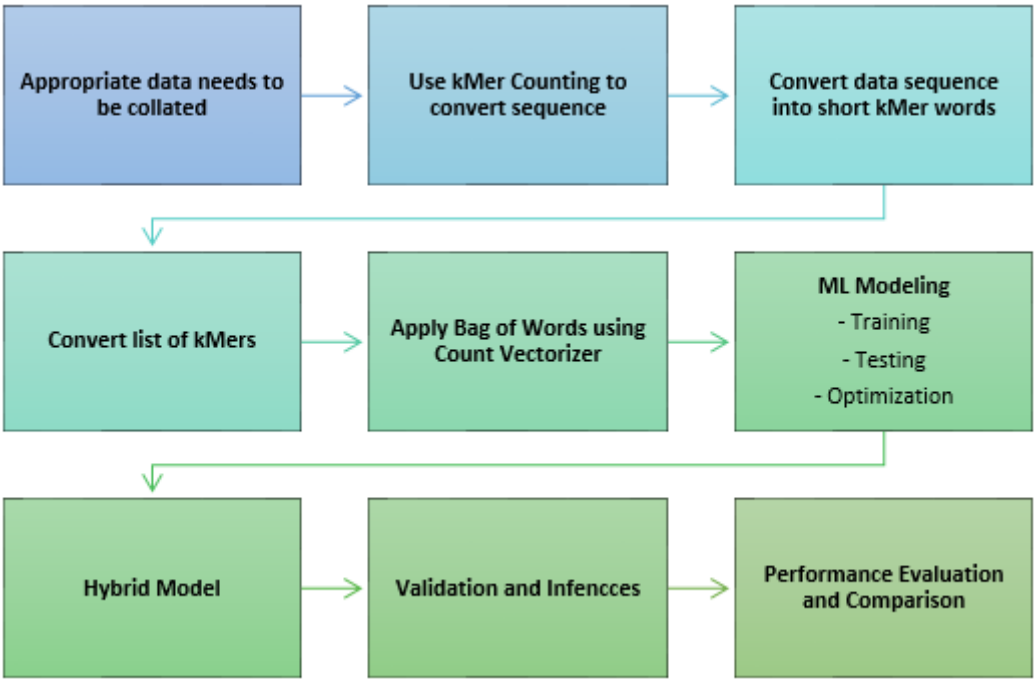


Figure 1. Workflow Architecture.

In the sub-sections below, we describe each of the modules of Figure 1 in detail, respectively.

2.2 Data Setting and Sample:

We undertook a detailed and investigative study of predictive models' derivation and validation using data collected for three different species - human, chimpanzee and dog.

The dataset consisted of genome strands from 4280 adult individuals for the human dataset, 1682 genome strands for chimpanzees and a comparatively more acute dataset for dogs, with 820 genome strands. The datasets' used in this study were categorized into three parts: A, B and C. 'A' dataset represented the human gene classification, 'B' represented the chimpanzee gene classification while 'C' took notice of the dog gene classification, categorizing them into seven different types of genes. In all the three datasets, there were two variables: DNA sequence and gene class.

Before we understand genome classification, it is important to understand what DNA sequences are, what are the major four chemical building blocks that make them, and how they tell the genetic information which helps machines to predict with the help of NLP.

DNA is a chemical molecule that consists of chromosomes, and makes up who we are in every sense, in accordance with inheritance. A DNA strand is made up of four chemical bases: adenine (A), cytosine (C), guanine (G), and thymine (T). For the two strands of DNA to zip together and form a hexical structure, A pairs with T, and C pairs with G. Each pair comprises a rung in the spiral DNA ladder. The order of these building blocks in a DNA molecule determines the genetic sequence and helps the machine to identify what the particular strand of DNA does in the body. These sequences make genes — the instructions for making specific proteins — and other genetic elements.

There were seven classes which the machine was trained upon including: G protein coupled receptors, Tyrosine kinase, Tyrosine phosphate. Synthetase, Synthase, ION channel and transcription factor.

1. G protein coupled receptors are membrane proteins that are used by cells to convert extracellular signals into responses inside the cell, including responses to hormones, neurotransmitters, and that of vision and taste [9].

2. Tyrosine kinases are important referees of this signal transduction process, leading to various cell functions such as programmed cell death [10].

3. Tyrosine phosphates remove phosphate groups from phosphorylated tyrosine residues, mostly working on proteins [11].

4. Synthetases take part in protein biosynthesis by catalyzing the attachment of a given amino acid to the 3' end of its cognate tRNA. They do this by forming an aminoacyl-adenylate of amino acid, which helps in transferring amino acid to tRNA [12].

5. Synthases work similarly to synthetases, except they do not utilize energy to perform these actions. [13]

6. ION channels provide passageways through which charged ions can cross into the plasma membrane. [14]

7. Transcription factors are proteins involved in the process of converting DNA into RNA. [15]

Figure 2 presents the DNA strands of synthases class. It gives a visualization of the DNA strands as they were presented in the hybrid model. All the nucleotides have been provided and marked with a color to indicate their presence inside the strand.

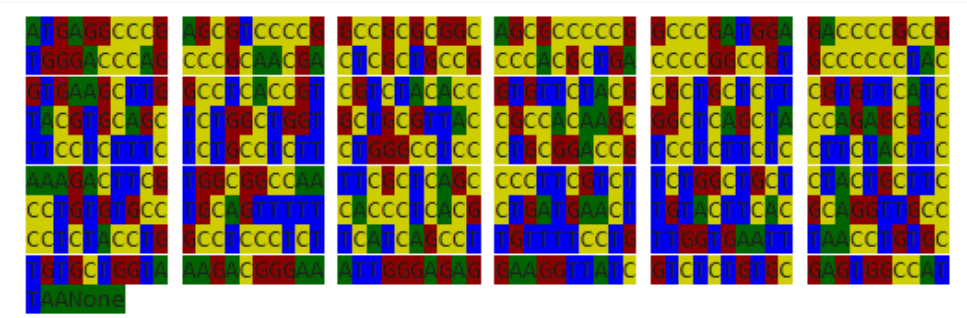


Figure 2. Pigmented Representation of the Nucleotides

Figure 3 presents the nucleotide composition for the strand illustrated in Figure 2. Inside the Figure 3, we observe that the 'C' nucleotide has the most chances to appear in the strand, which could also have been observed in the visual representation, by the yellow encoding being visually the most.

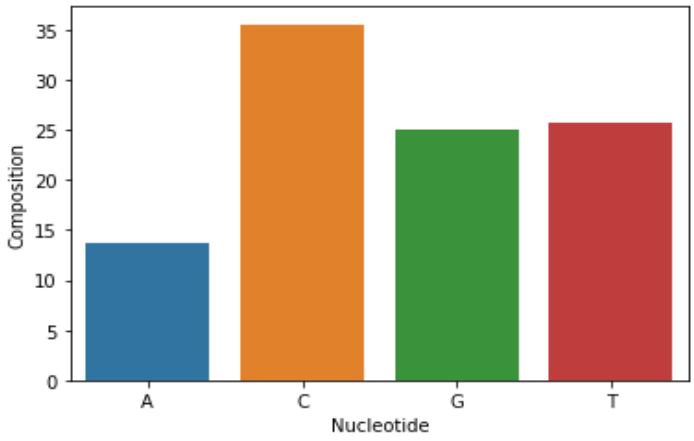


Figure 3. Bar Graph Depicting Composition of Nucleotides in The Strand

2.3 NLP Modelling

2.3.1 k-Mer Counting

The k-mer counting involves counting the number of substrings that have length k in a string S, or a set of strings, where k is a positive integer.

k-Mer counting was used in this study to break up the gene strands and convert longer strands into shorter ones, which helped the machine to understand the sequence better.

2.3.2 Convert data sequences: into short k-mer words

Once the kMer counting is complete, the model then converts these sequences into shorter kMer words by removing the blocks and simplifying the complex strand. The list of kMers is now converted into separate stands and the count vectorizer is then applied.

2.3.3 Apply Bag of Words

Bag of Words is a NLP model which helps in extracting features out of the text which can be helpful in machine learning algorithms. In bag of words, we get the occurrences of each word and construct the vocabulary for the corpus.

Bag of Words is used in this study to convert the data into a rawer data, so that it may be easily converted into numbers which can then be processed by the machine. CountVectorizer is an NLP based model provided by the scikit-learn library in Python. It is used to convert a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text, similarly using TFIDF value.

2.4 ML Modelling

Prediction models were built and validated on each of the three datasets separately.

For the development of an appropriate machine learning model based diagnostic aid for genes, a comparison of performance of various existing algorithms in the three datasets need to be presented. Preparation of the model is the most crucial step that provides the outline of the research. Steps that are included in the development of an appropriate model and tuning it for obtaining possibly the best result, are detailed below as already specified in the workflow diagram, Figure 1. The following sub-section describes both aspects.

2.4.1 Model Training: The dataset was split into two groups - training (75%) and testing (25%). The first included 75% of the dataset and was utilized for training the algorithm. The other group of 25% was used for testing, respectively.

We trained a model to identify strands with all of the seven classification using machine-learning tools. Model training was conducted using a diverse set of various machine learning algorithms. The algorithms that gave better results are the only ones presented here.

1. LOGISTIC REGRESSION
   Logistic regression may be a factual examination strategy to foresee a double result, such as yes or no, based on earlier perceptions of an information set. A calculated relapse show predicts a subordinate information variable by analyzing the relationship between one or more existing autonomous factors [16].

2. DECISION TREE
   A decision tree may be a type of supervised machine learning utilized to classify or make forecasts based on how a past set of questions were replied. The demonstration may be a frame of supervised learning, meaning that the demonstration is prepared and tried on a part of information that contains the required categorization [17].

3. RANDOM FOREST:

   The random forest is a machine learning algorithm that uses classification. It is made of many decision trees. It features randomness when trying to build every individual tree to give a higher accuracy than a singular one [18].

4. KNN

   kNN stands for k-Nearest Neighbours. It is a supervised learning algorithm. KNN works by finding the distances between a query and all the examples in the data. Once it has done so, it finds the specified number (k) of examples closest to the query and then predicts the more frequent one [19].

5. MULTINOMIAL NAIVE BAYES

   The Multinomial Naive Bayes algorithm uses a Bayesian learning approach mostly utilized in Natural Language Processing (NLP). The program tries to guess the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag's probability for a given sample and predicts based on a higher probability [20].

6. SVM

SVM works by mapping data into a feature space with higher dimensions so that data points can be categorized into the plane, even when the data is not linearly plottable. A separator between the categories is found to classify them into various types, then the separator is used for predictions. [21]

7. LINEAR SVC

   The objective of a Linear SVC (Support Vector Classifier) is to fit the data you provide into a graph, and give back a divider that divides the data onto the linear plane perfectly. Using this, the model can now make predictions.[22]

8. PASSIVE AGGRESSIVE

   Passive aggressive classifiers use a punish-reward system for regression programs, responding passively if the outcome is correct and aggressively if it is not.[23]

9. BERNOULLI NAIVE BAYES

   BernoulliNB uses the naive Bayes training and classification algorithms for data that is distributed according to multiple variable Bernoulli distributions, i.e., there may be multiple features inside the data sets however they are all binary.[24]

10. EXTRA TREE

    Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of multiple algorithms learning technique which uses and combines the results of several uncorrelated decision trees collected in a "forest" to give its prediction on the classification.[25]

11. BAGGING

    Bagging is a specific type of machine learning process that uses multiple learning techniques to better machine learning models. This technique uses various classified training sets where some features may be used multiple times between several training sets.[26]

12. GRADIENT BOOSTING

    Gradient boosting is a type of machine learning which betters and evolves the other models used. It uses the prediction that the next model, when used with former models, helps in lessening the overall prediction errors. The main aim is to set the target outcomes for the next model in order to get the best possible result. [27]

2.4.2 Optimization

Thus, in this study, we optimized certain algorithms employing hyperparameter optimization. Through this, it gave improved accuracy and other performance metrics in a much better standard than before. The optimization took place by employing certain factors in the algorithms and changing keys in the programs.

2.4.3 Model Testing

Model testing is referred to as the process in which the performance of a fully trained model is evaluated on a testing set. The testing set consists of a part of the dataset and is used to check the accuracy of the model by sending the input through the algorithm and comparing with real results. Evaluation is an integral part of the process of machine learning.

2.4.4 Performance Evaluation

In classification problems, success is measured utilizing the calculation of accuracy and precision of the model. In this study we have considered the following evaluation metrics:

1. Accuracy: Accuracy is defined as the percentage of correct predictions out of all the observations. A prediction can be said to be correct if it matches reality. Here, we have two conditions in which the Prediction matches with the Reality: True Positive and True Negative.
2. Precision: Precision is defined as the percentage of true positive cases versus all the cases where the prediction is true.
3. Recall: It considers true reality over prediction.
4. F1-score: F1 score can be defined as the measure of balance between precision and recall.
5. AUC: AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

2.5 Hybrid Modelling

A hybrid machine learning model employs more than one machine learning algorithm, and functions in such a way that one algorithm's output forms the basis of the second algorithm's input. By employing such factors, it helps in greatly reducing errors and exponentially increasing the accuracy rates.

In this study, we have employed stacking ensembles to build hybrid gene multinomial classifiers.

Out of 12 algorithms explained in Section 2.4.1, only 11 classifiers acted as a base classifier in constructing the ensemble model. In the final model construction, the classifier which negatively influenced the hybrid modelling, we dropped that classifier. Herein, we used a Logistic Regression classifier as a meta-learner. Also, we used a cross-validation (CV) of 5.

Cross-validation is a method to evaluate and cross-check ML models by training several ML models on subsets of the available input data and evaluating them on the remaining subset of the data, complementary to the one chosen earlier. CV is often used to detect overfitting, ie, failing to generalize a pattern.

2.6 Validation of the Built Model:

We calculated AUC- ROC curves, and precision-recall curves for the built hybrid model for the three species. AUC-ROC curve helps us visualize how well the machine learning classifier is performing. Although it is commonly used for binary classification, it can also be utilized to give a better validation in multi class classification problems. The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at several benchmark data values and, in its core, separates the true graph from the "noise" i.e. the unnecessary information that may cause overfitting. The Area Under the Curve (AUC) is often used as a summary of the ROC curve.

The higher the AUC, the better the performance of the model at differentiating between the various classes. The precision recall curve on the other hand shows the difference between the precision of recall for different data values to be able to compare the outputs differently.

3. Results

This section presents the different extensive experiments with the corresponding results in several subsections. The results of ML modelling and hybrid modelling are described for each of the three categories – human, chimpanzee and dog in the following subsections.

3.1 Machine Learning Modelling - HUMAN

Table 1 shows the quantitative results of the best performing ML models for humans.

Table 1. The best performing ML model.

ML Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score
Logistic Regression	92.60	93.0	93.0	0.92
Decision Tree	82.01	82.0	82.0	0.81
Random Forest	89.68	90.0	90.0	0.90
KNN (K=8)	77.08	84.0	77.0	0.78
Multinomial Naive Bayes	**97.35**	**97.0**	**97.0**	**0.97**
SVM	80.0	86.0	80.0	0.79
Linear SVC	92.05	92.0	92.0	0.92
Passive Aggressive	96.34	96.0	95.0	0.95
Bernoulli Naive Bayes	88.83	88.0	83.0	0.83
Extra Tree	81.0	81.0	81.0	0.81
Bagging	87.0	87.0	87.0	0.87
Gradient Boosting	83.47	87.0	83.0	0.83

All the classifiers demonstrate their respective best results for accuracy, precision, recall, and F1-score. Multinomial Naive Bayes, as shown in Table 1, beat all the classifiers in all the four-evaluation metrics. It generated a high accuracy of 97.35% for the gene classification, whereas KNN (K=8) gave the lowest accuracy of 77.08%.

Also, Multinomial Naive Bayes gave the highest precision, recall and F1-score, followed by Passive Aggressive, Logistic Regression, and Linear SVC. Extra Tree showed the lowest performance for precision. KNN showed the lowest performance for recall and F1-score. Furthermore, Passive Aggressive, Logistic Regression, Linear SVC, all had classification accuracy above 96%.

In terms of AUC, Multinomial Naive Bayes has the highest AUC, indicating that it is a better diagnostic predictor, with a 0.993 score, followed by Logistic Regression, Decision Trees, Random Forest, Naive Bayes, and SVM, which have 0.990, 0.988, 0.983, 0.981, and 0.973.

3.2 Hybrid Model - HUMAN

The summary of each model's capability of achieving the best accuracy from the proposed pipeline, with corresponding other performance evaluation measures, has been reported in Table 2.

Table 2. Hybrid Model Evaluations.

Accuracy (%)	Precision (%)	Recall (%)	F1 Score	AUC
98.08	98.0	98.0	0.98	1.00

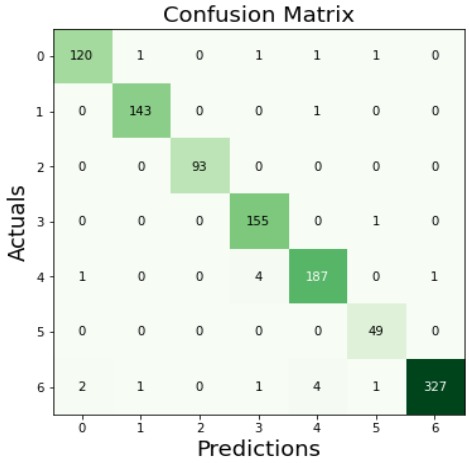
From the results of Table 2, we can comprehend that the ensemble model created for human species for the classification of seven gene types, gave the highly improved performance. Along with an accuracy of 98.08%, the built model resulted in a precision value of 98.0%, recall value of 98.0%, F1-score 0.98, and AUC of 1.00. Higher

accuracy does not mean greater machine learning model performance. It accounts for other performance evaluation measures as well, as described in Section 2.4.4.

A value of 98.0% for precision means that the constructed ensemble model returned more relevant results than irrelevant ones, and high recall value of 98.0% means that the model returned most of the relevant results (whether or not irrelevant ones were also returned). As the F1-score combines precision and recall, this implies that a value of 0.98 for human gene classification, a larger value of F1-score implies the constructed model is better than the non-hybrid models, as reported in Section 3.1.

Figure 4 (a) presents the confusion matrix for the hybrid model created. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. Herein the matrix, the diagonal (dark green shade) represents the correct prediction done for each of the seven classes. Through this confusion matrix, a series of performance evaluation measures was determined, as presented in Figure 4 (b). Precision, recall, F1-score for each of the seven classes for human species is presented. Overall, we got to achieve a good set of values, which implies that the build was able to predict each gene class efficiently.

The 2 classes (tyrosine kinase and transcription factor) achieved 100% precision and tyrosine kinase, along with ION Channel also received 100% recall. That means, the model was able to return 100% relevant results than the irrelevant ones.



(a)

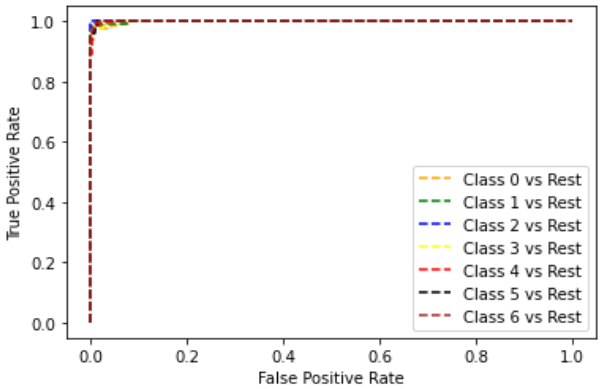
	precision	recall	f1-score
0	0.98	0.97	0.97
1	0.99	0.99	0.99
2	1.00	1.00	1.00
3	0.96	0.99	0.98
4	0.97	0.97	0.97
5	0.94	1.00	0.97
6	1.00	0.97	0.98

(b)

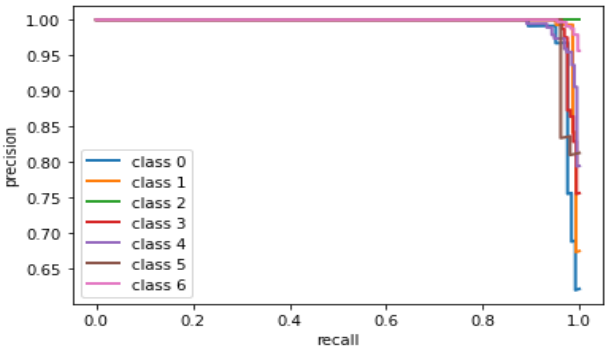
Figure 4. Confusion Matrix and Evaluation Values.

Figure 5 (a) represents the multiclass ROC curve for the built model. As the AUC-ROC curve is only for binary classification problems, we extended it to multiclass classification problems by using the 'One vs All' technique. In this technique, an AUC-ROC curve is calculated considering each label at a time and all the others are grouped as one label.

Thus, from Figure 5 (a), we can comprehend well that the AUC-ROC curve created for seven classes of genes fits well with an overall AUC score of 100%.



(a) Multiclass AUC-ROC curve for Human Data.



(b) Multiclass Precision-Recall Curve for Human Data
Figure 5. Plots for Human Hybrid Model.

Figure 5 (b) presents the precision-recall curve for created hybrid model in multiclass setting. As precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned, we saw in the results that we achieved a good balance in the precision and recall values.

The precision-recall curve shows the tradeoff between precision and recall for different thresholds. From the Figure 5 (b), a high area under the curve represented both high recall and high precision, where high precision related to a low false positive rate, and high recall related to a low false negative rate. High scores for both showed that the classifier is returning accurate results (high precision), as well as returning most of all positive results (high recall).

3.3 Machine Learning Modelling - CHIMPANZEE

Table 3. The best performing ML model.

ML Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score
Logistic Regression	84.81	88.0	85.0	0.88
Decision Tree	73.40	75.0	73.0	0.73
Random Forest	81.71	86.0	82.0	0.82
kNN (k=8)	67.22	78.0	67.0	0.68
Multinomial Naive Bayes	89.07	91.0	89.0	0.89
SVM	73.16	84.0	73.0	0.73

Linear SVC	89.54	92.0	90.0	0.90
Passive Aggressive	**89.55**	**92.0**	**90.0**	**0.90**
Bernoulli Naive Bayes	71.50	81.0	71.0	72.0
Extra Tree	71.02	72.0	71.0	0.71
Bagging	79.57	82.0	80.0	0.80
Gradient Boosting	77.21	81.0	77.0	0.77

All the classifiers demonstrate their respective best results for accuracy, precision, recall, and F1-score. Passive Aggressive, as shown in Table 3, beat all the classifiers in all the four-evaluation metrics. It generated a high accuracy of 89.55% for the gene classification, whereas Extra Tree gave the lowest accuracy of 77.08%.

Also, Passive Aggressive gave the highest precision, recall and F1-score, followed by Linear SVC and Multinomial Naive Bayes. Extra Tree showed the lowest performance for precision. On the other hand, KNN showed the lowest performance for recall and F1-score.

3.4 Hybrid Model - Chimpanzee

The summary of each model's capability of achieving the best accuracy from the proposed pipeline, with corresponding other performance evaluation measures, has been reported in Table 4.

Table 4. The Hybrid Model Evaluation

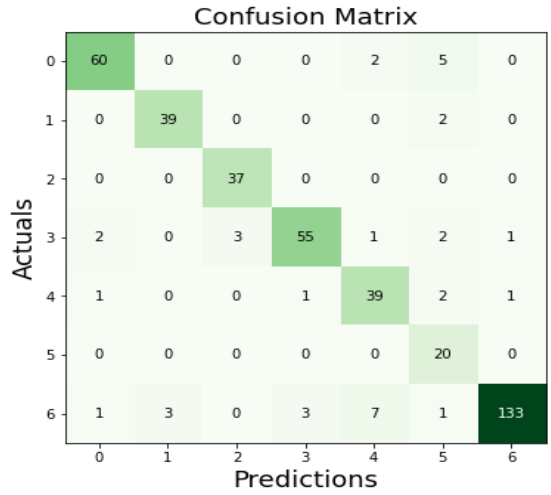
Accuracy (%)	Precision (%)	Recall (%)	F1 Score	AUC
90.73	92.0	91.0	0.91	0.990

From the results of Table 4, we can comprehend that the ensemble model created for chimpanzee species for the classification of seven gene types, gave a highly improved performance. Along with an accuracy of 90.73%, the built model resulted in a precision value of 92.0%, recall value of 91.0%, F1-score 0.91, and AUC of 0.990. Higher accuracy does not mean greater machine learning model performance. It accounts for other performance evaluation measures as well, as described in Section 2.4.4.

A value of 92.0% for precision means that the constructed ensemble model returned more relevant results than irrelevant ones, and high recall value of 91.0% means that the model returned most of the relevant results (whether or not irrelevant ones were also returned). As the F1-score combines precision and recall, this implies that a value of 0.91 for human gene classification, a larger value of F1-score implies the constructed model is better than the non-hybrid models, as reported in Section 3.1.

Figure 6 (a) presents the confusion matrix for the hybrid model created. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. Herein the matrix, the diagonal (dark green shade) represents the correct prediction done for each of the seven classes. Through this confusion matrix, a series of performance evaluation measures was determined, as presented in Figure 6 (b). Precision, recall, F1-score for each of the seven classes for chimpanzee species is presented. Overall, we got to achieve a good set of values, which implies that the build was able to predict each gene class efficiently.

The transcription factor class achieved 99% precision. That means, the model was able to return 99% relevant results than the irrelevant ones.



(a)

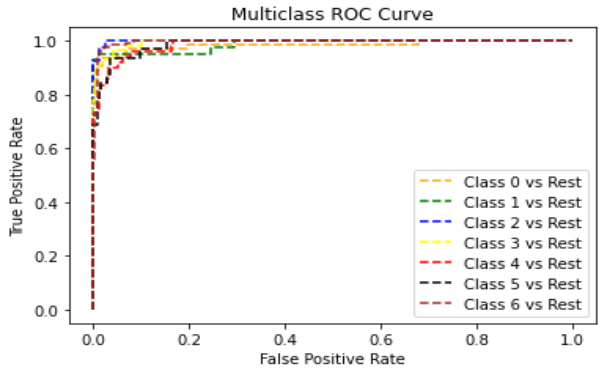
	precision	recall	f1-score
0	0.94	0.90	0.92
1	0.93	0.95	0.94
2	0.93	1.00	0.96
3	0.93	0.86	0.89
4	0.80	0.89	0.84
5	0.62	1.00	0.77
6	0.99	0.90	0.94

(b)

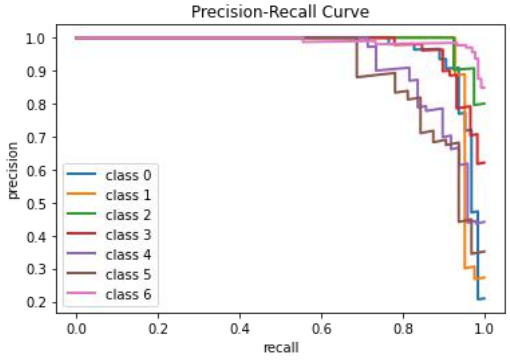
Figure 6. Confusion Matrix and Precision Values

Figure 7 (a) represents the multiclass ROC curve for the built model. As the AUC-ROC curve is only for binary classification problems, we extended it to multiclass classification problems by using the 'One vs All' technique. In this technique, an AUC-ROC curve is calculated considering each label at a time and all the others are grouped as one label.

Thus, from Figure 7 (a), we can comprehend well that the AUC-ROC curve created for seven classes of genes fits well with an overall AUC score of 99%.



(a)



(b)
Figure 7.

Figure 7 (b) presents the precision-recall curve for creating hybrid model in multiclass setting. As precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned, we saw in the results that we achieved a good balance in the precision and recall values.

The precision-recall curve shows the tradeoff between precision and recall for different thresholds. From the Figure75 (b), a high area under the curve represented both high recall and high precision, where high precision related to a low false positive rate, and high recall related to a low false negative rate. Comparatively higher scores for both showed that the classifier is returning accurate results (high precision), as well as returning most of all positive results (high recall).

3.5 Machine Learning Model - Dog

Table 5. The Best Performing ML Model

ML Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score
Logistic Regression	59.51	77.0	60.0	0.61
Decision Tree	55.12	58.0	55.0	0.56
Random Forest	59.02	76.0	59.0	0.60
kNN (k=8)	22.92	91.0	23.0	0.31
Multinomial Naive Bayes	**68.78**	**78.0**	**69.0**	**0.70**
SVM	48.29	80.0	48.0	0.53
Linear SVC	64.88	80.0	65.0	0.67
Passive Aggressive	64.39	79.0	64.0	0.66
Bernoulli Naive Bayes	40.97	78.0	41.0	0.47
Extra Tree	46.34	48.0	46.0	0.47
Bagging	57.07	65.0	57.0	0.58
Gradient Boosting	61.95	71.0	62.0	0.63

All the classifiers demonstrate their respective best results for accuracy, precision, recall, and F1-score. Multinomial Naive Bayes, as shown in Table 5, beat all the classifiers in all the four-evaluation metrics. It generated a high accuracy of 68.78% for the gene classification, whereas KNN (K=8) gave the lowest accuracy of 22.92%.

Also, Multinomial Naive Bayes gave the highest precision, recall and F1-score, followed by Passive Aggressive, Logistic Regression, and Linear SVC. Extra Tree showed the lowest performance for precision. KNN showed the lowest performance for recall and F1-score.

3.6 Hybrid Model - Dog

The summary of each model's capability of achieving the best accuracy from the proposed pipeline, with corresponding other performance evaluation measures, has been reported in Table 6.

Table 6. Hybrid Model Evaluation

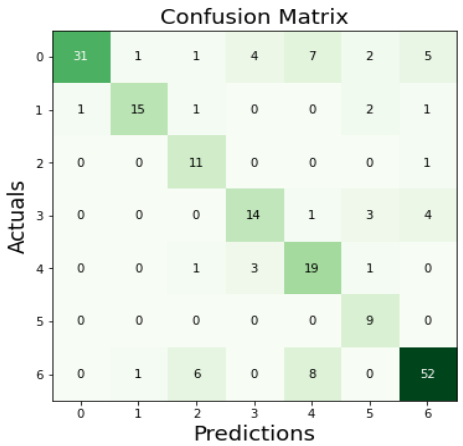
Accuracy (%)	Precision (%)	Recall (%)	F1 Score	AUC
73.65	79.0	74.0	0.74	93.80

From the results of Table 6, we can comprehend that the ensemble model created for dog species for the classification of seven gene types, gave a median performing performance. Along with an accuracy of 73.65%, the built model resulted in a precision value of 79.0%, recall value of 74.0%, F1-score 0.74, and AUC of 0.938. Higher accuracy does not mean greater machine learning model performance. It accounts for other performance evaluation measures as well, as described in Section 2.4.4.

A value of 79.0% for precision means that the constructed ensemble mode was majorly able to return more relevant results than irrelevant ones and recall value of 94.0% means that the model returned most of the relevant results (whether or not irrelevant ones were also returned). As the F1-score combines precision and recall, this implies that a value of 0.74 for human gene classification, a larger value of F1-score implies the constructed model is better than the non-hybrid models, as reported.

Figure 7 (a) presents the confusion matrix for the hybrid model created. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. Herein the matrix, the diagonal (dark green shade) represents the correct prediction done for each of the seven classes. Through this confusion matrix, a series of performance evaluation measures was determined, as presented in Figure 7 (b). Precision, recall, F1-score for each of the seven classes for dog species is presented. Overall, we got to achieve an average value.

The G coupled protein receptor class achieved 97% precision. That means, the model was able to return 97% relevant results than the irrelevant ones.



	precision	recall	f1-score
0	0.97	0.61	0.75
1	0.88	0.75	0.81
2	0.55	0.92	0.69
3	0.67	0.64	0.65
4	0.54	0.79	0.64
5	0.53	1.00	0.69
6	0.83	0.78	0.80

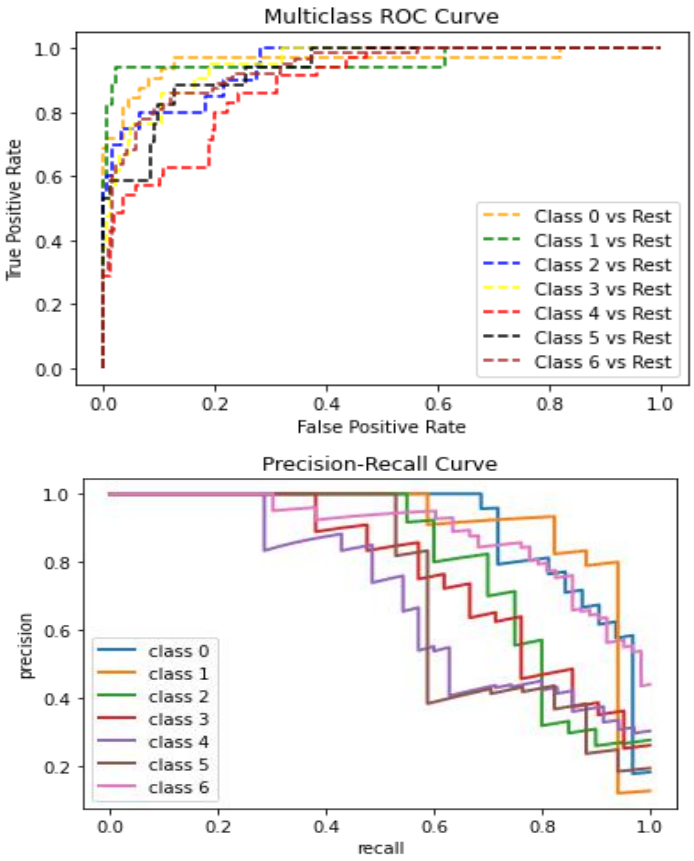


Figure 8.

A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labelled correctly.

4. Discussion

We compared and evaluated several classifiers against parameters such as accuracy, recall etc. used on three datasets to define the best fit for the hybrid model. When predicting gene categorization, we evaluated the calibration, optimization, and interrelation of machine learning-based prediction models. Machine learning techniques, when used in conjunction with other principles described in the learning healthcare systems approach, have the capacity to produce increasingly precise results. When adding novel models however, it is important to understand and note their uses in the medical world, their disadvantages over other models, and not just look at the accuracy they provide us with.

Different machine learning and NLP methodologies have been utilized by various studies. Gene classification has been used to detect diseases, especially ones such as cancer, diabetes, genetic background, healthcare and management. However, it has always been extremely hard to be able to choose the most suitable learning algorithm to solve specific problems that need solutions. In one of the studies, Christodoulou et al. conducted a systematic review where in 71 studies the performance of machine learning models did not significantly surpass the performance of logistic regression. Our results achieved an extremely high given accuracy and AUC score compared to most other studies inclusive of seven classification measures.

In the current study, therefore, we were able to select ensemble hybrid methods that had already been used in recent studies by other researchers to be able to give better performance levels in accuracy and AUC. The available methodologies that can be applied for multivariable classification problems are extremely wide ranging. Moreover, each one of these algorithms can be fine-tuned and optimized using various factors to give even better accuracy. Hence, even with very few algorithms there is an impossible number of possible combinations to give a very fine-tuned result, therefore, we tried to look at the most possible result combinations to be able to give a very good result with minimum tries. A limitation of this study is that we did not have enough available data entries available, especially for the dog species, which led to a lower accuracy for that study. Moreover, over one kind of database was used. With more data, the hybrid ensemble can be improved to even greater extents.

5. Conclusion

It is highly advisable to choose the better algorithms by applying various testing techniques upon them. Apart from choosing algorithms alone, hybrid/ensemble modelling methods are one to provide highly accurate results due to the types of variables and dependents in gene-classification related outcomes.

To conclude, our study looked at over twelve classifiers and stacked them to be able to classify the seven genome types. By looking and observing the difference in the DNA, scientists and researchers can continue to observe how different one species is from another and how differently the gene placement is done even if the gene performs the same function across several species. We were able to get a much better idea of how closely they were related to each other and classify them accordingly. An unfixed genetic code makes this an ever-changing field to research upon as genetic mutations continue to occur.

The opportunity of changing and bettering models arises more into the future as more and more data would become available over certain periods. Future research may include finding the pattern between nucleotides by machine learning to not only classify but predict the rest of the strand using DNA sequencing. However, such systems may bring in more problems as the reliability of the classifications may change a lot.

References

[1] https://www.ncbi.nlm.nih.gov/books/NBK12983/
[2] https://www.ncbi.nlm.nih.gov/books/NBK26821/
[3]https://www.sciencedirect.com/science/article/pii/S1532046417301685
[4]https://ieeexplore.ieee.org/abstract/document/9121260
[5]https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-92
[6]https://academic.oup.com/bioinformatics/article/32/7/1023/1743748?login=true#95428586
[7]https://www.hindawi.com/journals/cmmm/2021/1835056/
[8]https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/natural-language-processing
[9]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3967846/
[10]https://academic.oup.com/nar/article/43/22/10588/1803770
[11]https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/tyrosine-phosphorylation
[12]https://www.ncbi.nlm.nih.gov/books/NBK6026/#:~:text=Aminoacyl%2DtRNA%20synthetases%20(aaRS),amino%20acid%20to%20the%20tRNA.
[13]https://www.degruyter.com/document/doi/10.1515/bmc-2019-0001/html?lang=en
[14]https://www.nature.com/scitable/topicpage/ion-channel-14047658/
[15]https://www.ncbi.nlm.nih.gov/books/NBK26887/
[16]https://www.theseus.fi/bitstream/handle/10024/335764/Khadka_Birendra.pdf?sequence=2&isAllowed=y
[17]https://www.researchgate.net/publication/350386944_Classification_Based_on_Decision_Tree_Algorithm_for_Machine_Learning

[18]https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf
[19]https://ieeexplore.ieee.org/document/9065747
[20]https://scholar.google.co.in/scholar_url?url=https://perun.pmf.uns.ac.rs/radovanovic/dmsem/cd/install/Weka/doc/pubs/2004/KibriyaAI04-MultinomialNBRevisited.pdf&hl=en&sa=X&ei=1xQ_Yp73BZLeyQSQoL7oCQ&scisig=AAGBfm2WhED4yEScnh_pRvEluOD2XAILLA&oi=scholarr
[21]https://ieeexplore.ieee.org/document/708428
[22]https://www.researchgate.net/publication/221345963_Linear_Support_Vector_Machines
[23]https://jmlr.csail.mit.edu/papers/volume7/crammer06a/crammer06a.pdf
[24]https://ieeexplore.ieee.org/abstract/document/8776800/
[25]https://link.springer.com/article/10.1007/s10994-006-6226-1
[26]https://scholar.google.co.in/scholar_url?url=https://projecteuclid.org/journals/annals-of-statistics/volume-30/issue-4/Analyzing-bagging/10.1214/aos/1031689014.pdf&hl=en&sa=X&ei=tRY_YvDkPMOP6rQPsJSA6Ak&scisig=AAGBfm0GpWyUWm2swKA2EGSAlS4a8SN-9A&oi=scholarr
[27]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6511546/