Study Of Bacteria Classification Using DNA Sequence

- *By Parvathy Kurup*

Abstract:

In this study, snippets of specified length are analyzed using Raman spectroscopy which calculates the histogram of bases in the snippet. Each row of data contains a spectrum of histograms generated by repeated measurements of a sample, each row containing the output of all 286 histogram possibilities, which then has a bias spectrum (of random ATGC) subtracted from the results. The data (both train and test) also contains simulated measurement errors (of varying rates) for many of the samples, which makes the problem more challenging. The training set contains the spectrum of 10-mer histograms for each sample (per row). This analysis is to predict bacteria species based on repeated lossy measurements of DNA snippets. The dataset used for the following analysis contains 288 columns. When light interacts with molecules in a particular state of matter a vast majority of the photons are dispersed at the same energy as the incident photons. Raman spectroscopy uses this principle to understand and analyze the interaction after molecular vibration. This measures the inelastically scattered light from tissue that can identify native tissue and its biochemical constituents along with the changes associated with disease transformation. The genomic analysis following the principle mentioned is used to provide the data. Using various samples makes it possible to classify the compressed data into 10 different bacteria species. In this research, the analysis is applied to a dataset that follows the supervised nature of learning. The evaluation of the dataset is measured using categorization accuracy and confusion matrix. A brief study and analysis are executed using visualization tools to understand the non-linearity and nature of randomly generated data points within the data set. In this research, the classification is conducted using four-main shortlisted prediction algorithms to create a clear-cut comparison based on the performance evaluation obtained for each model. The algorithms are shortlisted based on the methodology, its compatibility with randomly generated data points, and computational strength. The aim of this study is to validate and present the power of machine learning in the field of genetic biotechnology. Some algorithms show better accuracy in gene identification than others. This also gives higher chances of avoiding the identification of false negative antibiotic resistance. Furthermore, it is to understand the concept of classification in real-time data and to showcase innovative technology.

Introduction:

The approach employs a genomic analysis technique that has some data compressed and data lost. This technique uses a block optical sequencing (BOS) method using surface-enhanced Raman spectroscopy (SERS).It gives a 10-mer length spectrum where the sequence of the DNA

will be lost but the BOC content is preserved. This technique is ideally suited for the genomic identification of bacteria clipping this genome into 10-mer lengths would provide enough DNA to form a single bacterium to cover the SERS pyramids. When light interacts with molecules in a particular state of matter a vast majority of the photons are dispersed at the same energy as the incident photons. Raman spectroscopy uses this principle to understand and analyze the interaction after molecular vibration. This measures the inelastically scattered light from tissue that can identify native tissue and its biochemical constituents along with the changes associated with disease transformation. The relationship between the incident and observational direction along with the refractive index of the sample plays a role in determining the accuracy between the incident and scattered intensity. The genomic analysis following the principle mentioned is used to provide the data. In fact, Blood infections contain very low counts of bacteria. Thus, a 10ml of blood would provide 100-fold more DNA than needed to place in a 10-mer on each pyramid. The DNA sequence obtained is broken into 10 base lengths. The BOC is measured for each 10-meter length, and they are put into bins corresponding to the fractional base spectrum (They are binned according to the percentage of A, T, G, C). Experimental noise is simulated by introducing random errors into sequence-specific FBC spectrum resulting in experimental FBC spectrum. Here, the spectrum from a purely random sequence (bias) is subtracted from the simulated experimental FBC spectrum, thus producing a deviation spectrum. The deviation spectrum of various DNA samples is analyzed by principal component analysis (PCA) and Machine Learning Algorithm. The classification model uses attributes where the values correspond to probability distribution from each FBC deviation spectrum. The initial approach is to employ 10-fold cross-validation and the train-test-split was a 90-10 split. The hyperparameter for FBC deviation spectra is *r* the number of pyramid tips for generating FBC spectra, *m* is the fractional error rate, and *s* is the number of FBC deviation spectra created per DNA sequence. E MLAs were tested based on the following categories:

1. Linear Machine Learning
2. Decision Tree Learning Algorithm
3. Random Forest Classification Algorithm
4. KNN Classifier Algorithms
5. Neural Network

The accuracies and confusion matrix for each MLA is the average of ten trials (n=10) for given r & m. To provide an easier approach, the labels will be encoded under integers and the aggregate maximum value is chosen finally as prediction outputs for testing in performance metrics.

The main challenge of the given dataset is the duplicated data generated on a random seed. Train and test distribution differ. If the score is computed without duplicating data in the validation set the classifier public score will reach higher than the cross-validation score. This observation suggests that there is a leak between the train and test data. The main challenge is to find the leak and exploit it. A third of the data and a quarter of the test data are duplicates. Several hundred training rows are duplicated in the test data. The sampling schemes yield a pre-sample of 200x1000 array for each sample. Finding the data transformation (and a suitable metric) such that the data points generated with similar seeds become close neighbors. The solution to data

transformation might involve making paired train and test points visible and by using a suitable model that matches the corresponding rows of train and test. To exploit the flawed random generator the priority would be to compute the greatest common divisor and integer representation. This bias is between 9.5 e-7 and 2.4e-2. The sum of all biases is 1. For each ATGC DNA sequence, a bias value is subtracted from the original FBC spectrum.

If the error rate is always a constant value of 1.0 then the output is a constant row. Using our prediction model, we find all pairs with appropriate distance metrics between the test sample and training below a given radius. This is determined by careful study of the correlation between the features by applying the appropriate metric to measure. The solution offered is theoretically written down from research and study. Since the actual test data points have no way of validating, in this study code the predictive model is executed only on the training dataset. The main goal of this study is to check if a 10-mer block of DNA information can be used by diagnostic instruments to correctly identify species and antibiotic-resistant genes with relativistic performance.

Literature Survey:

In [1], To understand the domain of genomic analysis and obtain the actual real-life problem of genetic disease and blood infections. This gives sufficient information to understand how this study can help to showcase innovative technology and help bring efficiency to biotechnology with the help of a computational algorithm.
In [2], To obtain a background in principles of Raman spectroscopy and how it adds to the formulation of the FBC spectrum. The relationship between incident direction and intensity is learned here. The nature of the dataset is produced from the optical instruments that follow this principle.
In [3], the basics of classification are Machine Learning algorithms to obtain an idea of what to employ for the multi-label classification problem. The various algorithms decided to shortlist involves Logistic Regression, Random Forest Ensemble, KNN Classification, and Neural Network
In [4], Due to the randomness in the created dataset, ensemble learning techniques were researched to employ on the dataset. This is believed to give a better result but as we are dealing with a large dataset it means using algorithms that are computationally costly.
In [5], This is to research and understand the different other variations of random forest classification where the splitting of the decision tree is under random order other than sequential.
In [6], ExtraTrees Classifier is one other extension to random forest classification where it does not employ greedy algorithm rather uses the entire dataset and splits values randomly to create decision trees of sub-groups.

In [7], Logistic Regression can work with multinomial or multiple classes hence this is to obtain the performance of a logistic regression model. Uses an LBFG Optimizer with l2 penalty as hyperparameters for large datasets with high dimensionality. This model makes use of probabilistic classification.
In [8], Understanding the working algorithm of Radius Neighbor Classifier which is an extension of KNN classifier and has the flexibility to use appropriate distance metrics.

In [9], Using TensorFlow to create deep learning models. The input node is set equivalent to the total number of features and the loss function is kept at sparse categorical entropy. This algorithm takes less computational time and memory space. TensorFlow architecture is easier to make and much highly efficient when dealing with biased data points. The performance should give the best results and the most accurate true value.
In [10], Deep learning uses Keras to further understand how to create stronger models. Understood how to obtain the best optimizers and sequentially ordered architecture. Involves the use of optimizers for categorization and loss functions compatible with the classification
In [11], To understand how diagnostics of genetic diseases and resistant genes work.

Reference	Objectives	Problem Statement	Methodology	Dataset	Algorithm	Advantage	Disadvantage	Performance Measure Value
1	To understand the domain of genomic analysis	Obtain background on the generation of BOC sequence from the DNA	Deviated FBC spectrum from the calculated bias	The main dataset	Deviated spectrum is obtained	Understand the domain of bioengineering	None	None
2	To understand Raman Spectroscopy	Raman spectroscopy	The relationship between incident direction and intensity	The main dataset	Refractive index of the incident and scattered deviation	To better understand the principles of this analysis	None	None
3	To understand the algorithms used for classification	Obtain the notion of machine learning algorithms	The various algorithms are shortlisted to Linear, DT,LDA and NN	The main dataset	The Algorithms shortlisted	Helps to understand the best algorithm	None	Confusion Matrix
4	To study an ensemble of random forests	Use of random forest and decision trees	Random forest is used for the classification of bacterial species	The main dataset is	Greedy Algorithm	It gives better performance	Computationally Costly	F1-Score
5	To study extremely randomized forest	The data points are under randomly generated order	Random order other than sequential	main dataset	Employs random order decision	Compatible with the random	Does not give a good	Cross-Validation metric

					tree algorithm	nature of the data points	public lb score	
6	Effects of Extra Trees Classifier	Effective use of classification	Extreme randomized where splitting of decision trees operates in random order	main training dataset	Extension to Random Forest Classification algorithm	Less computational cost	High bias and variance	Cross-Validation metric
7	Logistic regression model on a training dataset	Can work with multinomial or multiple classes	Uses an LBFG Optimizer with l2 penalty	Main dataset	Probabilistic Classification	It is compatible with classification tasks involving multiple classes	May not give the best performance result	Accuracy score or confusion matrix
8	Radius Neighbors Classifier model	Performance score using the Radius Classifier model	Acts as an extension to the KNN algorithm but this model can make use of distance metric	The training dataset when decomposed to its higher integer values can be used under this model	The distance metric with appropriate radial hyperparameter tuning can provide good results	It can make use of a radial measure with appropriate distance metric	Takes more memory usage and computation.	Accuracy score and confusion matrix
9	Deep learning architecture using TensorFlow	NN for a classification task	The input node is set equivalent to the total number of features and the loss function is kept at sparse categorical entropy	The training dataset is standardized	Architecture is easier and highly efficient	Less computational time and memory usage	None	Accuracy score and confusion matrix

10	Deep learning architecture with Keras	To obtain the best optimizers and sequentially ordered architecture	Possible to access a wider number of optimizers and loss function	Main dataset	Sequential models for classification.	Makes the architecture easier and more efficient	None	None
11	To obtain better background in the practical application of diagnostics	This is to understand how this study can be used in real-life application	The rapid separation of bacteria from blood infections	Using the training dataset and test set	Infected blood contains more count of bacterium which gives more amount of variance in its DNA sequence	To understand how this study can be an innovative approach to real-life application	None	None

Implementation:

The process of simulating experimented data points and transforming them into a spectrum by subtracting them with a formulated bias gives the initial dataset that is dealt with. It has 200000 rows followed by 288 columns. As this is a multi-label classification problem, the list of names was found to be 10 unique targets under the same resistant gene. The initial exploratory data analysis was conducted for a better understanding of the data points. Random graphical visualization of deviated RBC spectrum gave the necessary points to conclude the duplicated data present in the dataset. Fig 1.1 depicts the randomly picked FBC spectrum (eg: A0T3G1C6, A4T3G0C3… etc), and the scatter points shows the frequency of histogram bases under each of the 10-target label. Due to the duplicated values, the overview of the plotted figures shows overcrowding of points. Next, visualization tools were implemented to obtain the count plot graph of all 10 target names (that is *Bacteroides fragilis, Campylobacter jejuni, Enterococcus hirae, Escherichia coli, Escherichia fergusonii, Klebsiella pneumoniae, Salmonella enterica, Staphylococcus aureus, Streptococcus pneumoniae,* and *Streptococcus pyogenes*). Two additional species (that is *Klebsiella aerogenes* and *Mycobacterium tuberculosis*) were only used for testing data.

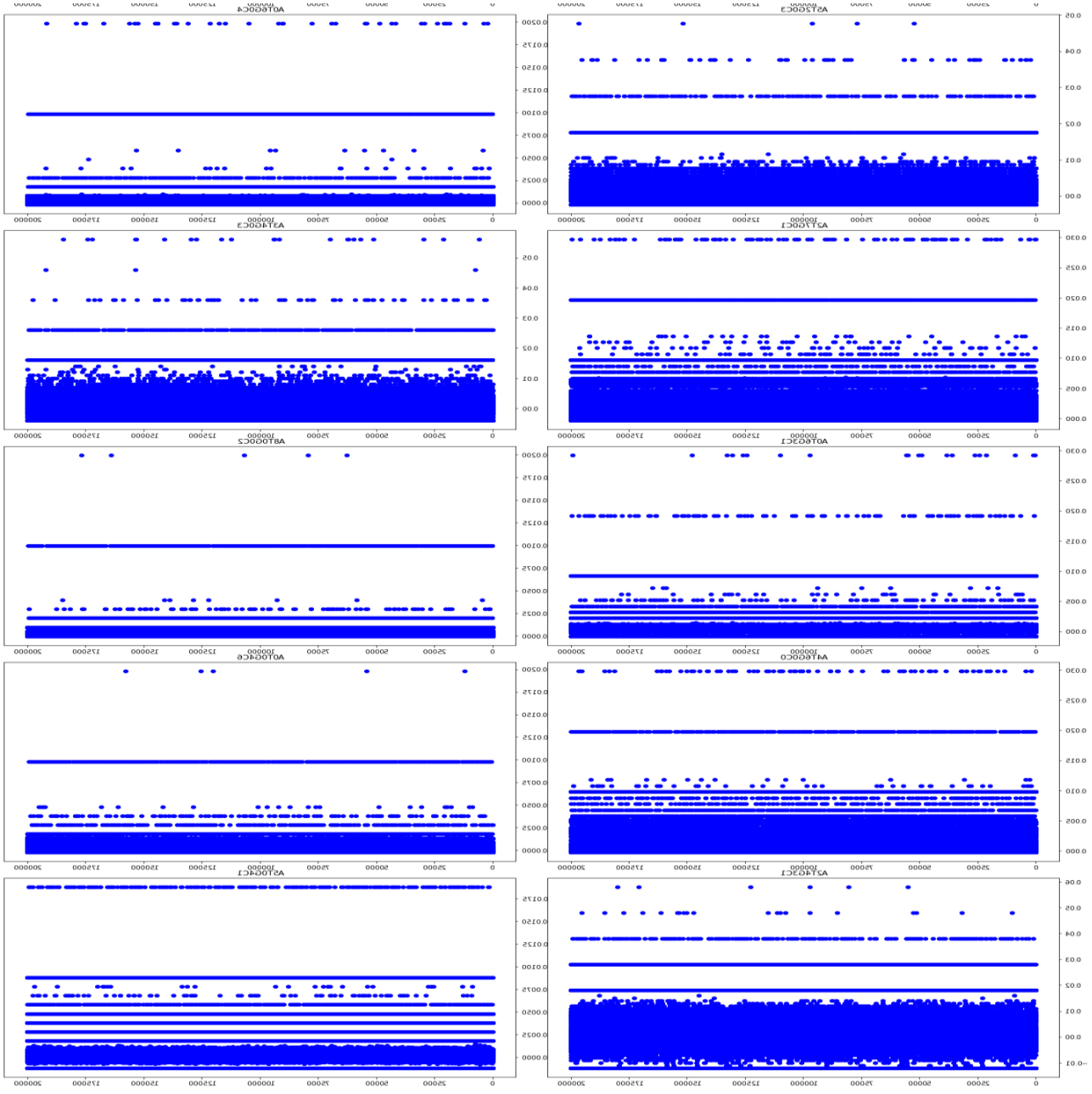


Fig 1.1

To understand the generation of the deviation spectrum under each of the target labels, Fig 1.2 depicts random target bloodstream infection for the generated spectrum of histogram bases. Since there are 286 different possibilities the graph shows the acute difference in 10-mer frequency deviation from bias against the histogram distributions (of A,T,G,C). The optical instrument reads the BOC of each DNA of 10-mer bound to a limited number of SERS pyramids. Since the size of each read is known, the base proportions are converted into specific integers

for each snippet. BOC reads as seen in Fig 1.1 is shown as AwTxGyCz where 0 ≤ w,x,y,z ≤ k, and w+x+y+z = k.

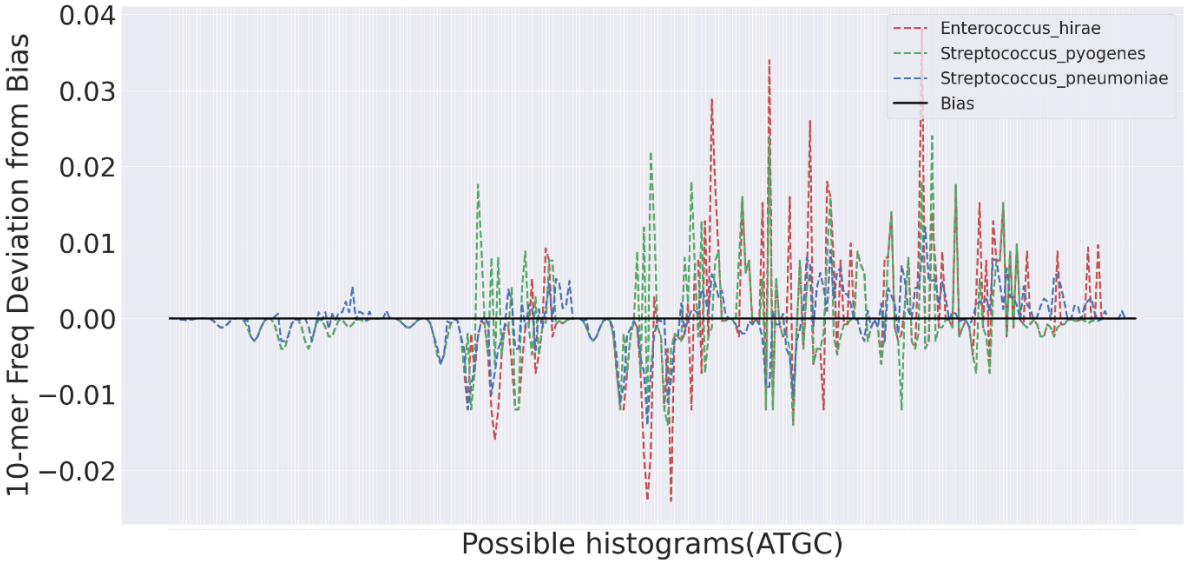


Fig 1.2

In order to create the FBC spectrum obtained in the above visual of Fig 1.2 each sequence is decomposed into every possible 10-mer length block. The blocks are then evaluated under assigned corresponding bins to produce the final FBC spectrum. Once all the blocks are sorted into their bins for the given sequence, each bin count is divided by the total 10-mer count of all 286 bins for that DNA sequence to get the specific sequence of probabilistic distribution or the FBC spectrum. As mutations are bound to be present in bacteria, no gene would show a perfectly identical FBC spectrum, hence random experimental errors are produced to the spectrum by involving a chosen error rate. Furthermore, in order to enhance the dataset and prevent overfitting, the FBC spectrum is subtracted with a totally random ATGC giving a spectrum of deviation in random order nature. A purely random spectrum or the bias spectrum in this case can be calculated as:

Bias Vector = 10! / (w!x!y!z!*4^10) (1)

Using this information, FBC deviation spectra created for each specific sequence of both training and test set are exploited due to the randomly generated values to yield higher accuracy.

Architecture Diagram:

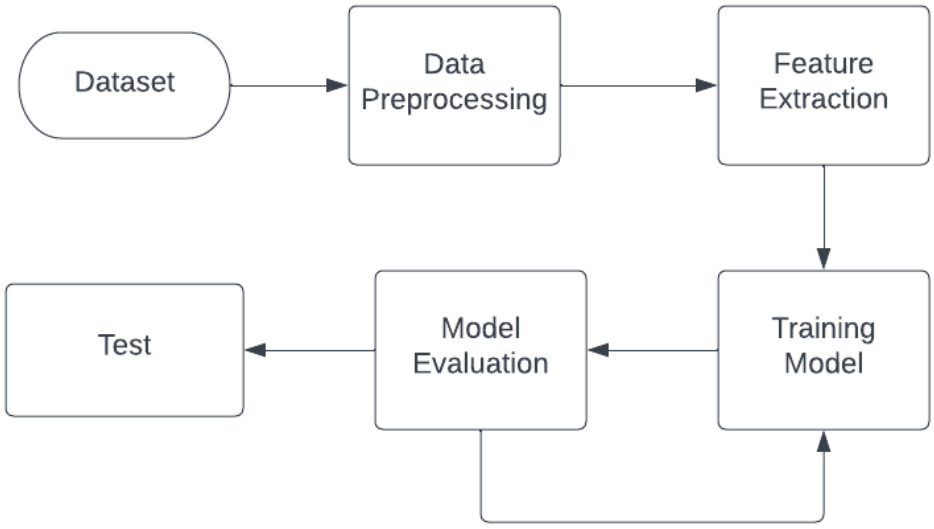


Fig 1.3 Overall diagram for executing classification models

Algorithm:

To truly understand the generated spectrum-specific data points a new methodology was formulated that involved the use of Principal Component Analysis and feature transformation. PCA is a useful algorithm for the reduction of dimensions while keeping the trends and information of high variance. New variables are produced equivalent to the number of dimensions present in the dataset. They are constructed such that they show a linear combination or mixture of the initial variables. The new variables produced are always uncorrelated to one another. Each new variable or component will contain the maximum possible information that further decreases as the next components are produced. Organizing this way allows the reduction of dimensions. This does not produce the best results for classification, but it plays a vital role in the pairing of training and test data points to eliminate the randomly generated duplicate points. Here, PCA was used to transform data points into just two components so as to depict the scatter plot to show the relationship between the training data points and the testing data points.

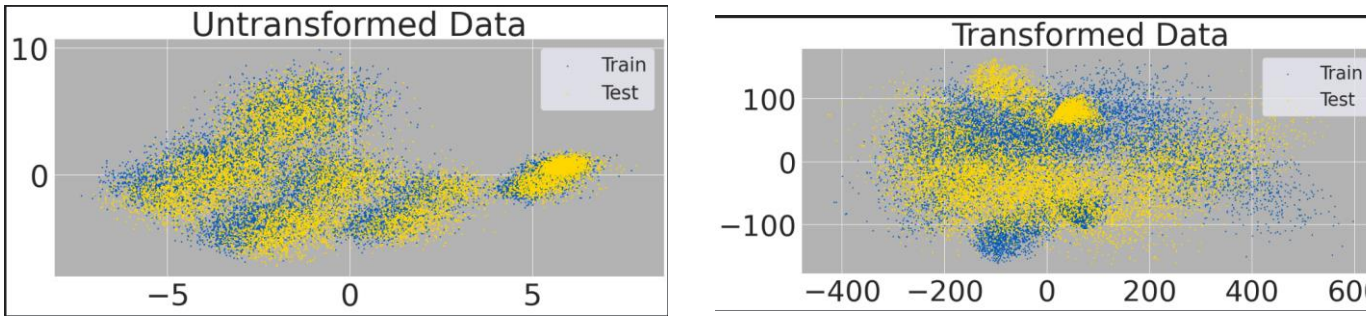


Fig 1.4(a,b)

As seen in Fig 1.4, the untransformed graph shows the unpaired training data points clustered due to the duplicate values. To solve the issue, randomly generated values under the common seed are exploited. First, the bias vector is obtained by computationally executing equation (1). On running a comparison of the rows with respect to their common divisors, the duplicates are removed and made into a new data frame. The data frame is transformed by converting it to a 2D array so as to reduce the dimensionality for the final scatter plot. The result is Fig 1.4, a transformed graph that clearly depicts the paired training and test data points.

The training data was randomly split in a 9:1 ratio and run through four main algorithms based on the categories shortlisted for the problem. The categories involved, linear machine learning algorithms, decision tree learning algorithms, ensemble classifiers, K-nearest neighbor classifiers, and TensorFlow neural network.

This model uses the probabilistic classification of datasets. Since there are multiple features in this dataset the multinomial nature of logistic regression makes it compatible with the dataset. The dataset used here is of larger memory hence the optimizer is set at LBFG (Limited-memory Broyden Fletcher Goldfarb Shanno) and the penalty at l2 with maximum iterations over 1000.

The Radius Neighbors Classifier is a machine learning algorithm that is an extension of the k-nearest neighbor's algorithm. The predictions are made here using all examples within a given radius of a new example instead of taking the k-closest neighbors. Here the radius taken was to be 18 and the model was made to compute additional observations such as the unique predictions, frequencies, samples, and predicted samples.

Neural Network implemented under TensorFlow used the Sequential modeling API. The input layer takes up to 128 nodes, followed by three hidden layers and an output layer that has 10 nodes for each of the 10 target labels. For comparison of accuracy, two models were created to distinguish between the label-encoded targets. While model-1 followed the integer encoded targets for prediction, model-2 follows target labels encoded under a binary matrix. Both models use the ReLu activation function with the output activation function kept as SoftMax. The dropout layer is added after the hidden layer to prevent overfitting. Fig 1.5 shows the rough architecture of the artificial neural network used.

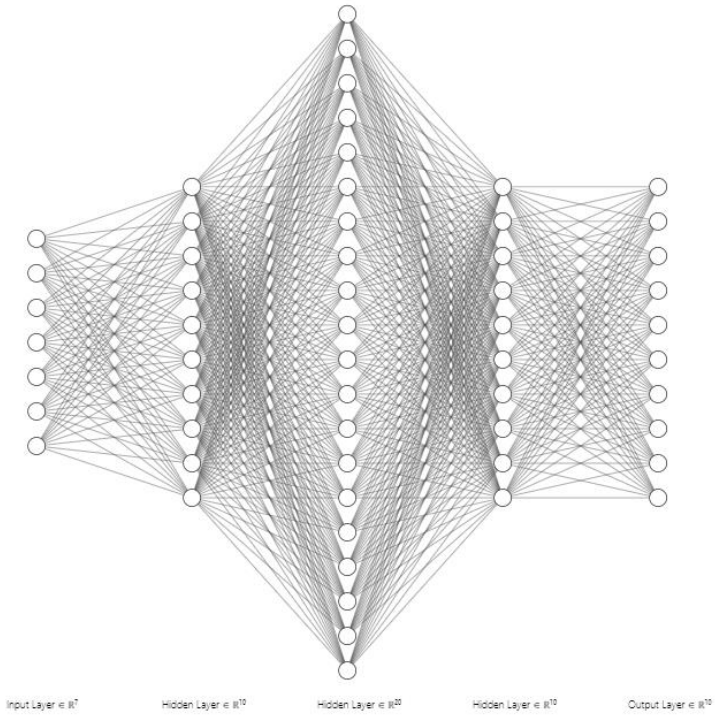


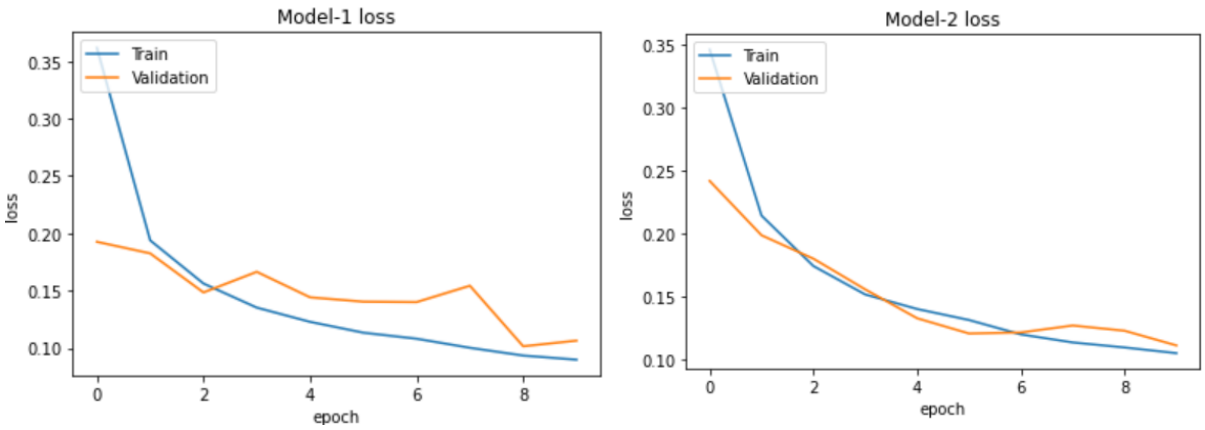
Fig 1.5 *Rough architecture of Neural Network that starts with 128 input neurons, three hidden layers [256,512,256], and 10 output neurons (since the problem is a multi-label classification*

ExtraTrees classifier is an extension of the random forest and decision tree algorithm; the model performs an extremely randomized forest where the splitting of decision trees operates in random order. It is an ensemble ML approach that trains several decision trees and aggregates the results of each sub-grouped tree to produce the output final prediction. Extra Trees do the process of splitting the values to create child nodes randomly hence decreasing computational cost and increasing time efficiency. However, If the dataset does not undergo pre-modeling steps such as feature selection, then the model might end up giving high bias and variance. Therefore, we employ an additional ensemble technique called Stacking along with the above-mentioned classifier. Here the predictions of the classifier are stacked, and they get used as new features to train the meta-classifier which makes the final prediction. Here the meta-classifier and the level-one classifiers and are all kept as the same ET model. The stacking is done for three Extra Trees models. This is a type of ensemble classifier that has the capability with our dataset to outperform all other algorithms.

Results and Discussion:

In doing the comparative analysis of the accuracy and performance of algorithms on the dataset. Principal component analysis was not found to be a reliable algorithm alone for purpose of classification. Logistic Regression on 1000 iterations gave an accuracy score of 87.6% for training data. Radius Neighbors Classifier takes more memory space, but the accuracy score

increased up to 95.98% on the training data.
The experimented ExtraTrees classifier with stacking served to produce the best result of 99.62% accuracy on training data and 96.18% on the test data. The two models of Neural Networks – one working under integer encoded labels (Model-1) and the other under binary matrix labels (Model 2). Model 2 showed a performance of accuracy of 96.48% with a decreased loss function.



Conclusion:

This study gave good exposure to understanding how classification algorithms work and gave a thought-provoking experience to the power of machine learning. Biotechnology and genomic analysis is an ever-growing field and bringing new technology to reduce time, cost, and energy for procedures and techniques that normally take many hours is a boon. Due to the large dataset and the random error implemented within the values, the algorithms used for this study did take up higher computational costs than the average. Managing memory usage was also a challenge faced while using the dataset. The leak between the training dataset and test set also was something that didn't entirely help with understanding if the model was overfitting. Nevertheless, the results yielded were substantially good and the overall experience was notable.

References:

1. Wood RL, Jensen T, Wadsworth C, Clement M, Nagpal P and Pitt WG (2020) Analysis of Identification Method for Bacterial Species and Antibiotic Resistance Genes Using Optical Data From DNA Oligomers. Front. Microbiol. 11:257. doi: 10.3389/fmicb.2020.00257
2. Wang, D., He, P., Wang, Z., Li, G., Majed, N. and Gu, A.Z., 2020. Advances in single cell Raman spectroscopy technologies for biological and environmental applications. *Current opinion in biotechnology*, *64*, pp.218-229.
3. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," in *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153-1176, Secondquarter 2016, doi: 10.1109/COMST.2015.2494502.
4. Shaik, A.B., Srinivasan, S. (2019). A Brief Survey on Random Forest Ensembles in Classification Model. In: Bhattacharyya, S., Hassanien, A., Gupta, D., Khanna, A., Pan, I. (eds) International

Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, vol 56. Springer, Singapore.

5. Extremely randomized trees Pierre Geurts · Damien Ernst · Louis Wehenkel
6. B. Dhananjay, N. P. Venkatesh, A. Bhardwaj and J. Sivaraman, "Cardiac signals classification based on Extra Trees model," *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2021, pp. 402-406, doi: 10.1109/SPIN52536.2021.9565992.
7. Wright, Raymond E. "Logistic regression." (1995).
8. Y. Lee, "Handwritten Digit Recognition Using K Nearest-Neighbor, Radial-Basis Function, and Backpropagation Neural Networks," in *Neural Computation*, vol. 3, no. 3, pp. 440-449, Sept. 1991, doi: 10.1162/neco.1991.3.3.440.
9. F. Ertam and G. Aydın, "Data classification with deep learning using Tensorflow," *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 755-758, doi: 10.1109/UBMK.2017.8093521.
10. Géron, Aurélien. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.
11. Alizadeh, M., Wood, R. L., Buchanan, C. M., Bledsoe, C. G., Wood, M. E., McClellan, D. S., et al. (2017). Rapid separation of bacteria from blood - Chemical aspects. Colloids Surf. B Biointerfaces 154, 365–372. doi: 10.1016/j. colsurfb.2017.03.027