SMU Data Science Review

Volume 2 | Number 2 Article 13

2019

Predicting Premature Birth Risk with cfRNA

Jason Lin
Southern Methodist University, jasonlin388@live.com

Jonathan Marin
Southern Methodist University, marinj@mail.smu.edu

John Santerre
Southern Methodist University, john.santerre@gmail.com

Predicting Premature Birth Risk with cfRNA

Jason Lin, Jonathan Marin, and John Santerre Ph.D.

Southern Methodist University, Dallas TX 75205, USA

Abstract. Identifying which genes are early indicators for preterm births using cell-free ribonucleic acid (cfRNA) from non-invasive blood tests provided by pregnant women can improve prenatal care. Currently, there are no medical tests for early detection of preterm birth risk in routine checkups for pregnant women. Recent studies have shown potential genes that can predict preterm birth. Machine learning techniques are utilized to see if the Area Under the Curve (AUC) can be improved upon when evaluating the prediction accuracy for chosen genes sequences and concentrations. Using cell-free RNA data from non-invasive blood tests in conjunction with machine learning, we improve upon the current methodology in an effort to identify and provide evidence between gene expression data and preterm birth. In our analysis, the model accuracy is improved using cfRNA Sequence Counts by expanding the feature space in which we have increased model AUC from 81% to 100%. These results are intended to provide additional evidence of model validity as an early indicator of preterm birth.

Keywords preterm delivery, predicting, cfRNA, RNA, data science.

1 Introduction

Preterm birth is the leading cause of death in infants where 15 million babies are born prematurely before 37 weeks every year. Rates of preterm birth range from 5% to 18% across 184 countries.[1] Earlier weeks of delivery may incur complications to the infant and may affect viability. Early detection of preterm births allows doctors to provide early treatment plans to help infants cross over to later weeks of viability. Countless studies and analysis have been done on the biology of fetal development though there has not been a proposed test that is both accurate and easy to implement. Present medical methods, ultrasounds and the last menstrual period, are imprecise and are easily miscalculated and misinterpreted [1]. These tests can only measure the gestational age of the baby and not the risks of preterm birth.

The most recent Stanford study by Ngo, T. and Moufarrej, M. et al., utilized cfRNA data from pregnant women, proposed two models. The first model was a random forest regression model that utilizes 8 cfRNA to predict gestational age with high accuracy. This data set only contained a Denmark cohort

[1] "Preterm Birth." World Health Organization. February 19, 2018. Accessed July 07, 2019. https://www.who.int/news-room/fact-sheets/detail/preterm-birth

of Caucasian women who all delivered full term. Ngo, T. and Moufarrej, M. et al. then attempted to use another cohort containing preterm data as a test set to attempt to predict time to delivery for women who delivered preterm. Ngo, T. and Moufarrej, M. et al. stated this was not a good result in trying to predict preterm birth.

Ngo, T. and Moufarrej, M. et al. also used a hierarchical clustering model with 7 cfRNA to predict 2 months in advance of preterm delivery with 81% AUC, area under the curve [1]. However, some drawbacks of these models are that further tests need to be done on a larger population to ensure its accuracy and result since the number of observations used is 15 for the hierarchical cluster model [2]. Note that for the regression and clustering models, Ngo, T. and Moufarrej, M. et al. used two different sets of data to predict gestational age and preterm birth risk.

Hierarchical clustering model analysis is the main problem of interest since it has proposed genes for early indicators of preterm birth risk. The hierarchical model concluded an AUC of 81% based on the validation test [2], however, this is based solely on one model type. Throughout the Ngo, T. and Moufarrej, M. et. al. paper, there are no indications of other models used or tested outside of hierarchical clustering to predict preterm birth. Therefore, the study will expand on this by utilizing a wider feature space and implement machine learning techniques to redetermine what genes are early indicators for preterm births.

We developed a higher AUC in three ways. First, Extra Tree classifiers are used to determine gene importance and number of genes to use in model development. Second, we are using grid search on the given model algorithm in order to determine model parameters that outputs the highest AUC. Third, we are using a stratified shuffle split method to help combat over-fitting of the model and to avoid splitting the data set into train and test since the data is small. Further analysis must be done on the reasoning of why no attempt was made to deal with the small sample issue i.e. 15 observations containing University of Pennsylvania cohort and 38 observations containing University of Pennsylvania cohort (PENN) and University of Alabama at Birmingham cohort (UAB) [1]. Exploration of different validation techniques are done to see if the model provides unbiased results.

There are many different data mining, machine learning, and statistical techniques that can be implemented to increase AUC with different costs associated with each. Each model has its own assumptions, and therefore analysis must be done to ensure to not violate any that may bias the results.

Analysis of the RNA sequence count data using only the 15 women from the PENN cohort showed some overlapping with the 40 genes deemed as potential indicators in the previous study [1] when feature importance was used to determine what genes have predictive power. Further exploratory analysis also shows that this is a balance data set where there are equal numbers of preterm and full term births. When a decision tree and random forest was attempted, the AUC of the models shows to be 100% when 33 genes are used. This could be indication of over-fitting and therefore the number of features should be paired

down. When scrutinizing the top seven genes considered to be of high importance in the feature importance methodology, it shows RPS6P25, UBE2S, FAM84B, HIST2H3C, KLHL14, TFIP11,and HIST2H2BE. The result here is completely different from previous study where they found CLCN3, DAPP1, MAP3K7CL, MOB1B, PPBP, RAB27B, AND RGS18 as the most predictive genes [1].

Analysis of the RNA concentration using the 38 women from UAB and PENN showed completely different results compared to the RNA sequence count data. The AUC of the model when using decision tree is 80%, random forest is 77%, and K nearest neighbors (KNN) is 67% when using 49 genes in the model. The small data issue may be occurring in that the accuracy result may not be robust, and therefore bootstrapping and bagging may need to be used to increase consistency of results. However, apart from the stated issue, the top seven genes that are considered of high importance are RPL23AP7, S100A8, TBC1D15, OAZ1, POLE2, RAB27B, and DAPP1. Which as stated earlier is completely different from the previous result as shown in Ngo, T. and Moufarrej, M. et al, except for gene RAB27B [1].

Therefore, when looking at the results it seems to show that the two different data sets have different indicators what are considered important genes for preterm risk birth. When different models are attempted, the accuracy also fluctuates. However, this result is confounded by the small sample issue and different measures of gene expression. For a clearer result, attempts should be made to mitigate the small data issue to introduce more robust and consistent results by implementing stratified shuffle split cross validation. This cross validation method allows repeated train and test on subsets of the data in order to obtain an ensembled AUC.

2 Data Collection

The original study made exclusions pertaining to certain women because of medical issues outside the scope of preterm birth or sample issues. The supplemental material provided by Ngo and Moufarrej et. al. describes the medical characteristics of the two cohorts University of Alabama at Birmingham (UAB) and University of Pennsylvania (PENN) and the collection and quantification of the blood samples.

We focused on two data sets provided by Ngo and Moufarref et.al. The first data set contained RNA sequence counts data from only the PENN cohort. The second data set contained RNA concentration counts for both the UAB and Penn cohorts. The UAB sample of 26 women all have history of preterm delivery, however, three were excluded because the blood sample taken was nine weeks prior to delivery [3]. The PENN sample contains 15 women that were studied where one had preeclampsia [3]. In the RNA sequence count data, only the PENN cohort was used in the hierarchical study [1,3] conducted by Ngo and Moufarref et.al. The RNA concentration data set, used as another potential data set which is not used by Ngo and Moufarrej et.al., is the combination of UAB

and PENN. For the PENN cohort, two women were dropped from the group by Ngo and Moufarref [3].

These exclusions are important when it comes to the interpretation of the results. These exogenous variables introduces bias since the medical history of the women may not be representative of the population. Therefore, variables such as this should be kept in mind in order to parse out outside factors influencing what genes are considered important in the model.

These blood samples are used to determine the genes and their given levels in the sample. In both the UAB and PENN cohorts, the study had only one blood sample taken before birth [3]. The method used to measure genes in the blood sample is RT-qPCR, quantitative reverse transcription.

As an overview of the RT-qPCR described by ThermoFischer Scientific, the samples of messenger RNA (mRNA) or total RNA is first converted to complementary DNA (cDNA) using reverse transcriptase. Then the cDNA is used as the template for the polymerase chain reaction (PCR) in order to have a measurable amount of DNA to quantify what the gene is expressed. A one-step or two-step process can be used where in one-step the reverse transcriptase and PCR are done in one tube, while two-step process uses two separate tubes. Both processes have advantages and disadvantages in the accuracy of the results. This general process and the advantages and disadvantages of RT-qCR can be found from the ThermoFisher Scienitific website [5] [6].[2] The Ngo and Moufarrej et. al. study uses the one-step RT-qPCR process to measure the different genes in the sample using total RNA, with also different methodologies for sequencing for each different study [3].

The genes measured in the data set encodes for many different parts of human body. However, the original study does not provide any specific dictionary to what the genes encode. Therefore, in order to cross reference to what the genes encodes for in the human body, the human genome website [7] is used as reference for the gene dictionary.[3] There is also no unit of measure for the gene concentration for the given patient since the number represents the florescence in reference to the control/non- reactive sample florescence [8] .

In reference to the study, it mainly focuses on the genes for placenta, immune, and liver since previous studies have shown that placenta and liver gene concentrations have correlation to pregnancies [3]. Koh and Pan et. al., studied how the genes for placenta and liver do have varying concentrations in pregnant women depending on what trimester the blood sample is collected, but also how these genes are specific to fetal development because of the temporal trend and high concentrations during pregnancy [9]. However, other genes, apart from liver, immune, and placenta, are also measured, as found in the data set.

[2] ThermoFisher Scientific. https://www.thermofisher.com/us/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/basic-principles-rt-qpcr.html. Last accessed 23 Mar 2019.

[3] Human Genome Resources at NCBI. https://www.ncbi.nlm.nih.gov/genome/guide/human/. Last accessed 23 Mar 2019.

In the data set for gene concentration, it contains the 23 women from the UAB cohort and the 13 women from the PENN cohort. The methodology described above is the process used to collect and quantify the blood sample for analysis. However, in the RNA sequence count data set that contains 15 women from the PENN cohort, further processing is done on the sample. Of the 15 sample collected, the RNA is sequenced and mapped to the human genome using STAR aligner, and further quantification is used to determine count using an algorithm called htseq-count.

Blood samples were examined from pregnant women in order to distinguish women at risk of spontaneously delivering preterm or not [1]. These blood samples contain cellular and cell-free RNA that are specific to the organ to be measured which in this case is the placenta, immune system, and fetal liver [4]. RNA transcript profiling using micro-arrays and RNA sequencing takes measurements of thousands of protein-coding and non-coding genes[4].

For the RNA Sequence Count data set containing only the PENN cohort, we used Ngo and Moufarref et.al spreadsheet that was provided and had to prep the data by pivoting it for a data science algorithms by having the features as columns and the last column being the target variable. The data was previously normalized by the Ngo and Moufarref et.al study. The original data set received had the targets listed in the column names itself which had to be corrected. No other feature engineering, interaction variables, nor external data were used in our analysis for better comparison to the Ngo and Moufarref et.al study.

For the RNA Concentration data set containing the PENN and UAB cohort, we had to normalize the data set. Features that were missing data were also dropped from the feature space. No feature engineering, interaction variables, nor external data were created for this data set. However, we created two target variables for evaluation of classifier and regression machine learning models from what was provided.

Figure 1 shows summary statistics of the two cohorts. The UAB cohort consisted of 15 women that were at risk of preterm birth because of symptoms shown, however seven women delivered at full term and the other eight at preterm. Samples for the PENN cohort were only collected once and at the time of delivery. The cohort for UAB had 26 pregnant women in which only five had delivered preterm spontaneously and eighteen have delivered to full term. [1] The women from both UAB and PENN cohorts were all African-American. The study did not have any Hispanic samples and there are no Caucasian samples that delivered preterm from these cohorts.

Table 1. Summary Statistics of Cohorts Used [1]

	Pennsylvania (n=7) Full Term	Pennsylvania (n=8) Pre Term	Alabama (n=18) Full Term	Alabama (n-5) Pre Term
Age(years, mean)	23.7	23.0	25.28	25.8
BMI(kg/m2, mean)	31.9	25.1	28.6	33.0
Ethnicity - Hispanic(Counts)	0	0	0	0
Ethnicity - Caucasian(Counts)	0	0	0	0
Ethnicity - African-American(Counts)	7	8	17	5
Gestational Age at Delivery (weeks, mean)	39.4	26.4	38.7	30.6

3 Summary of Hierarchical Study

The Ngo and Moufarrej et. al. study investigated which genes have explanatory power in determining if the birth is preterm or not. The study utilized RNA sequenced data and found that there are 38 potential genes that can differentiate between preterm and full term [1]. The methodology used to determine the 38 potential genes are the following statistical tests: Exact Test, Likelihood Ratio Test, and Quasi-likelihood F test [1]. After this was determined, a False Discovery rate test and Hedges g was used as a statistical test to determine if there is statistical significance in the genes predicting preterm and not. The study also used a combination analysis to validate the UAB cohort. The combination analysis the study implements is where a combination of genes of size 3 is used to validate the cohort. In total, 13 combinations are used in this validation. The validation result shows an area under the curve of 0.81 and 0.86 for UAB and Denmark respectively.

4 Feature and Sample Methodology

4.1 Feature Importance and Threshold Importance

Using the two data sets provided by Ngo and Moufarrej et. al. study, feature importance has been attempted on both. Feature importance can help determine what features have better explanatory power. Features can be added into the decision tree to see how much a reduction has occurred in the given criterion, which in this case is the Gini importance [14]. By graphing what features have caused greater reduction in the given criterion, it helps select the top features to use in order to limit the risk of over-fitting and run time.

Extra Tree Classifiers i.e. Extremely Randomized Trees is used as our main methodology in determining most predictive variables from the feature space. Extra Tree Classifiers is similar to random forest except Extra Trees does random selection of features and splits at the node when computing the decision tree and the random forest [19]. This is less computationally expensive.

The Gini impurity measure is used to gauge the importance of each feature. The impurity reduction caused by the inclusion of the feature is the method used to order feature importance. Therefore, the higher the number the more important the feature. The following figure 1 and figure 2 shows the illustration of initial features ordered by importance.

However, because of how large the feature space is, not all features can be used in the models or over-fitting will occur. Therefore, a threshold is used to reduce the number of features. The feature selection is used to set the threshold and see how many features pass this threshold [20].[4] Three different thresholds are implemented to see which genes are selected: Features that exceed the mean

[4] "Sklearn.feature selection.SelectFromModel." Scikit. Accessed July 19, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.feature selection.SelectFromModel.html.

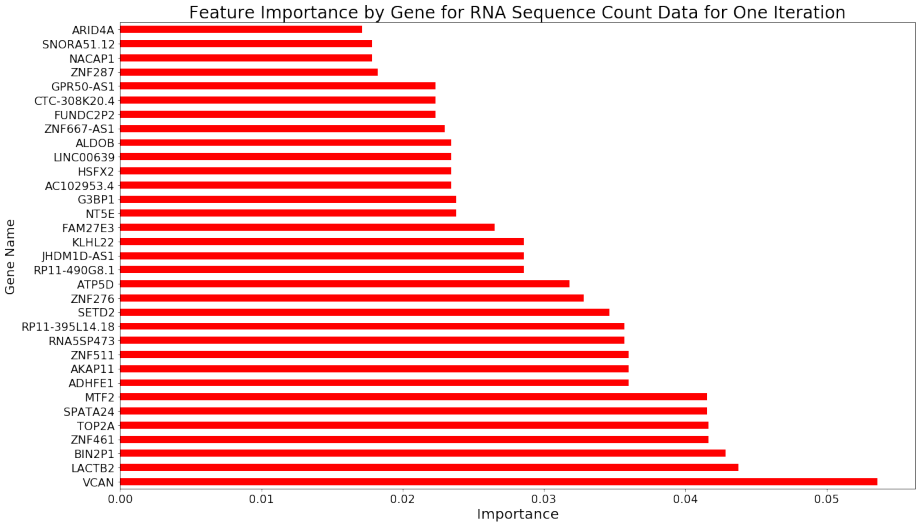

Fig. 1. Feature Importance of Genes in RNA Sequence Count Data

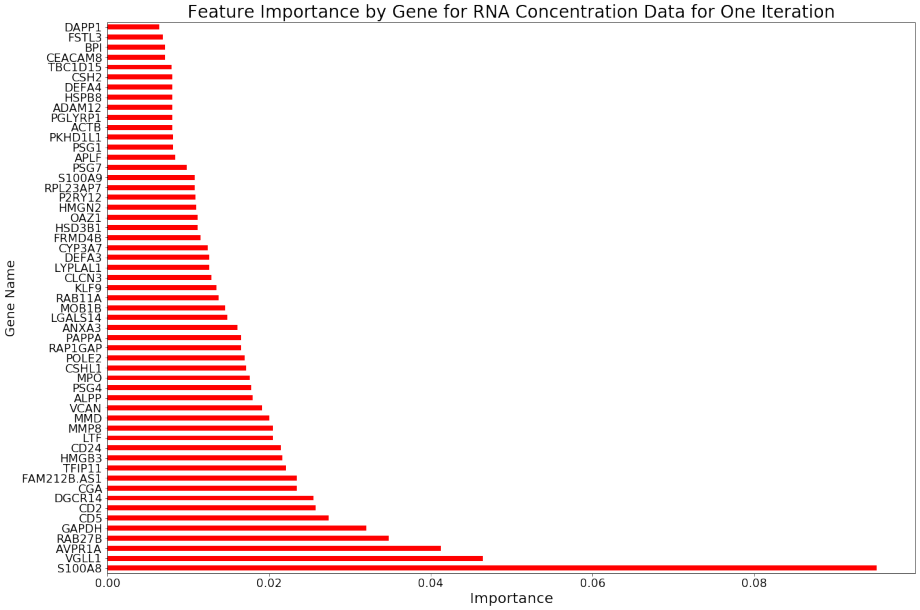

Fig. 2. Feature Importance of Genes in RNA Concentration Data

of feature importance, 1.5 times the mean of feature importance, and 0.5 times the mean of feature importance.

Feature selection resulted an issue was found when the algorithm is run successively. Because of the small sample size, it was found that the feature list is unstable by producing different important features on subsequent runs. Many features are constantly kept and dropped with each run. Therefore the feature selection is ran 100,000 times and we track how many times each gene is selected during a 100,000 run.

Figure 3 shows the average number of selected genes and list of genes. On average 33 genes are selected for the 100,000 runs, and the following are the top 33 frequency count for the RNA sequence count data. Figure 4 shows the average number of selected genes and list of genes. On average 49 genes are selected for the 100,000 runs, and the following are the top 49 frequency count for the RNA sequence count data.

The threshold used for figure 3 is the mean of the feature importance. When comparing to the 1.5 mean and the 0.5 mean, it was found that the feature list did not change much in genes selected. Also the number of genes selected did not fluctuate out of what is expected. The threshold used for figure 4 is 0.5 mean of the feature importance.There is more fluctuation in the ordering of the features between the different thresholds indicating less stability in gene selection. However, since there is more samples in this data set, it is decided to allow more features to be selected. A plot was made once the feature list is

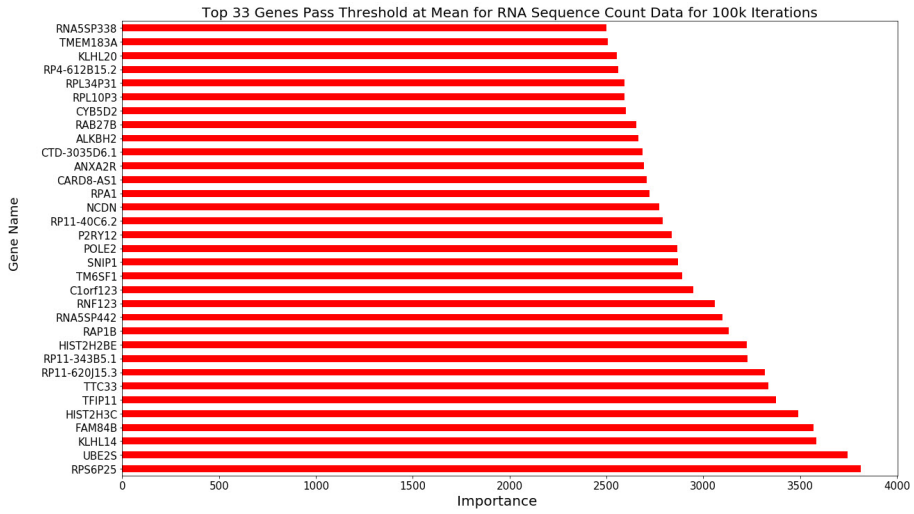

Fig. 3. Feature Importance of Genes in RNA Sequence Count Data for 100,000 Run

selected. The plot is used to observe how well the genes are at differentiating between preterm and full term births. If a clear separation is seen for the given gene, then the gene may have high predictive power. Figure 5 shows the plots of the 33 genes selected for the RNA sequence count data.

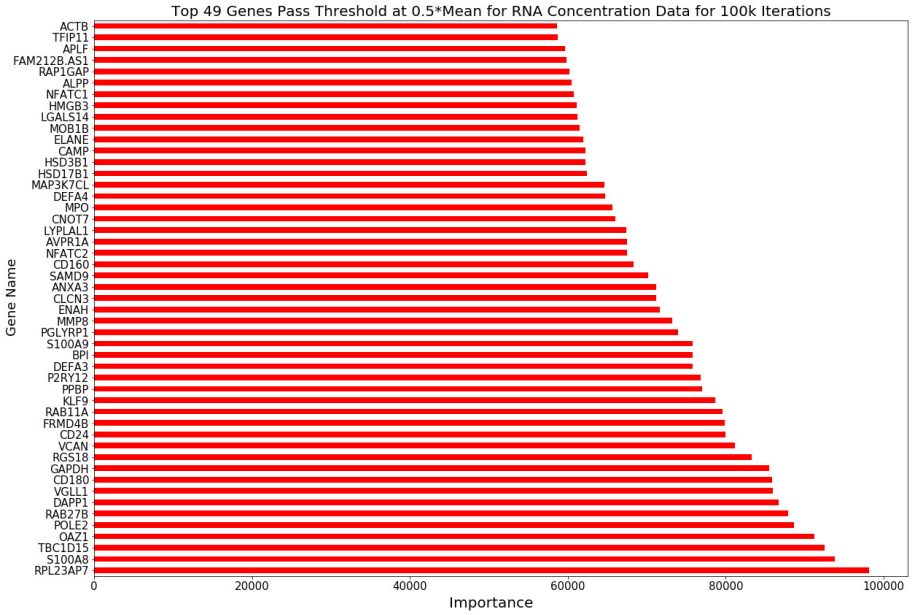

Fig. 4. Feature Importance of Genes in RNA Concentration Data for 100,000 Run

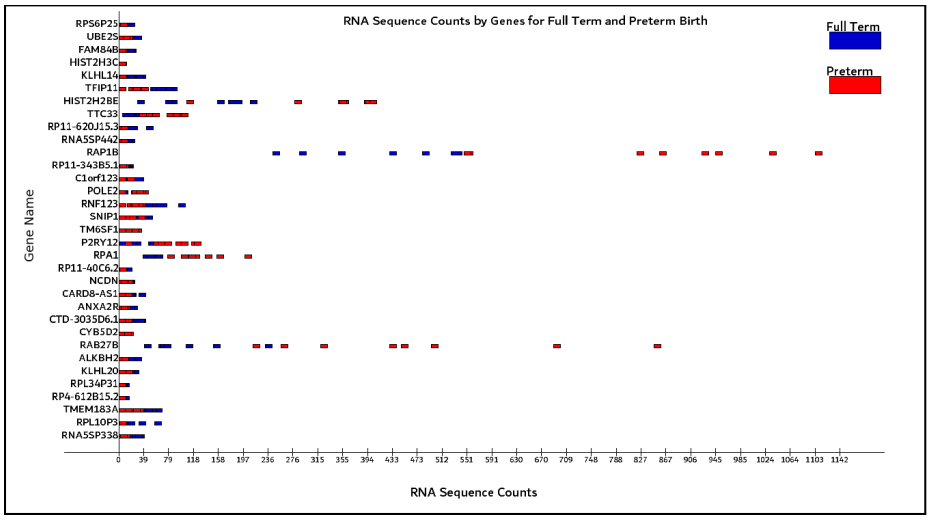

Fig. 5. Plot of Top 33 Genes for RNA Sequence Counts

As seen in figure 5, many of the genes have power to clearly differentiate between preterm and full term births. For example, gene RAP1B shows clear

differentiation between full term and preterm births for the 15 women sample. In RAP1B, it shows that women with higher counts of this gene, pass a certain threshold, are more likely to have preterm births compared to women with lower counts of the gene. The plot helps to see the gene counts and birth type relationship, which can be used as the thresholds for medical tests in later implementation.

4.2 Training/Test Splits and Stratified Shuffle Split Validation

In the original study, the training data set for RNA sequence count was one entire cohort (Pennsylvania) while the testing set was a different cohort (University of Alabama Birmingham). The traditional training/test split was used in the validation of the model. The original study had never mentioned the use of any cross validation method for predicting preterm births. The study also never used the RNA concentration data in there analysis. Therefore, no validation technique was used for the RNA concentration data set.

Since there is only 15 observations for the RNA sequence count and 36 observations for the RNA concentration data set, the issue of over-fitting may occur. With over-fitting the results may not be applicable outside of the test case, and therefore care must be taken when interpreting results. In this study, it was decided to use cross validation, specifically, stratified shuffle split cross validation was used [21].[5] Cross validation is one the many techniques used to combat over-fitting. The method allows for the test set to be truly unseen, and the model can be validated many times on random sets to see the average fit of the model.

In stratified shuffle split, a percentage of the training data set is used to create the testing data set. The parameter used in this study is 20% of the training data. The number of iterations to test the model is also specified. The parameter used in this study is 10 splits. Therefore, for each of the 10 iterations, 20% of the data is placed in the test set at random. However, in the stratified shuffle split, the percentage of the class variable in training and test set is maintained. In essence, no training and test set would be of one class type i.e. all full term or preterm. The average score is taken once the model is trained and tested ten times.

The score used in this study for the fit of the model is the AUC (Area Under the Curve). The AUC represents the area underneath the ROC (Receiver Operating Characteristics) [22] [23].[6] The AUC measures how well does the model clearly separates the 1's as 1's and 0's as 0's. The range that the AUC can take on is between 0 and 1, where 0 the model is classifying all observations wrong and 1 is the model can clearly differentiate between classes. An AUC of

[5] "Sklearn.model selection.StratifiedShuffleSplit." Scikit. Accessed July 19, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.model selection.StratifiedShuffleSplit.html.

[6] "Sklearn.metrics.roc auc score." Scikit. Accessed July 19, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.rocaucscore.html sklearn .metrics.rocaucscore.

0.5 indicates the model cannot differentiate between the two classes. Therefore, the higher the score the better the model is at differentiating between classes i.e. preterm and full term births.

5 Classification Algorithms

Once the data is cleaned and the features are selected, the following sections are the background information on the algorithms and methods used to test if the genes selected are potential indicators for preterm birth risk. Many classification algorithms are used to see if the indicators are robust across all methodologies, and if not what assumptions are made to reach this different conclusion.

5.1 Decision Trees and Random Forest

The decision tree is typically a model that sequentially asks questions to lead us to a certain result and can be applied to both categorical and numerical data. This is easily interpreted and represented in the model. This is a simple model that does not require any assumptions and is easily interpreted with short computational time [15].[7] The number of questions that are asked is the depth of the tree, and therefore the more questions asked can lead to a more accurate result. This model is also a non-parametric model meaning no information is needed about the statistical distribution of the data. However, the drawbacks of this method are what is considered an optimal decision at each question i.e. node since the optimal decision at the given node does not guarantee an optimal result [15]. Also, in order to obtain an optimal result many questions are asked, which can lead to over-fitting. In decision trees the number of nodes can be capped to a certain amount, however this leads to issues with error due to bias since it is likely the optimal answer may not be reach [15].

Therefore, random forest is used to combat these issues that the decision tree has by creating multiple decision trees. Note random forest is also a non-parametric model meaning no information is needed about the distribution of the data in order to run the model. By creating multiple decision trees, random forest is able to reduce variance by training on different samples of the data [15]. Since random forest is an ensemble method, each decision tree casts a vote for a given outcome and the majority vote is the final result. Because of this random forest can use all features in the data set by having each tree ask different questions and the combined result of each of these trees would take account for these different features [15]. The main assumption of this model is that the features must have low correlation. This prevents decision trees from influencing one another biasing the result. As seen here, random forest and decision trees are used to help see what indicators predictive power in preterm birth through the questions asked at each node.

[7] Liberman, N. (2019). Decision Trees and Random Forests. [online] Towards Data Science. Available at: https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991 [Accessed 4 Jun. 2019].

5.2 Clustering Methodologies

Clustering methods are pursued because of the ease of implementation, but also there is no data size requirement to run this methodology. These models are also non-parametric models meaning no information is needed on the distribution of the data. Clustering is also able to give a measure of accuracy for the ability of the gene to differentiate between preterm and full term births. This is accomplished by iterative filtering the data set to one gene expression and then seeing where the points cluster for preterm and full term births. The cluster method used in this paper is K Nearest Neighbors (KNN).

In KNN, the clustering is accomplished by first placing a random point in the field. Then a distance metric is calculated using either Euclidean, Manhattan, or Minkowski of the random point from the data point. After the distances are calculate, the distances are then ordered from smallest to largest, and the first k numbers are considered one cluster [17].[8] In this cluster, majority vote is made to determine what is the outcome i.e. the cluster is full term or preterm births. The main drawback of this method is that computational time may increase as the data set becomes larger. In KNN, the whole data set is used in training the model. Therefore, the larger the data set the longer the computational time and storage needed to run and implement the model

6 Results

6.1 RNA Sequence Count Data Set

The feature importance results for the RNA sequence count data set is shown in Figure 3 above and Figure 6. The mean threshold is used as stated earlier. As seen in the figures, the highest count is considered of high explanatory power in the model. The top seven genes considered to be of high importance are RPS6P25, UBE2S, FAM84B, HIST2H3C, KLHL14, TFIP11, and HIST2H2BE. The result here is completely different from previous study where they found CLCN3, DAPP1, MAP3K7CL, MOB1B, PPBP, RAB27B, AND RGS18 as the main genes [1].One thing to note, the figure shows that the gene with the highest count is 4000 out of a 100,000 run. This further supports the instability in the feature importance when using Extra Trees classifier.

[8] Seif, G. (2019). The 5 Clustering Algorithms Data Scientists Need to Know. [online] Towards Data Science. Available at: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 [Accessed 4 Jun. 2019].

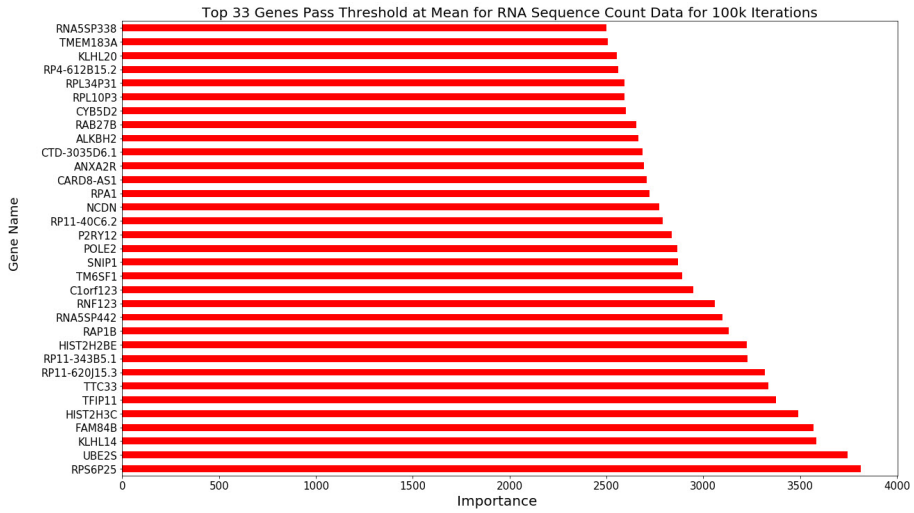

Fig. 6. Feature Importance of Genes in RNA Sequence Count Data for 100,000 Run

The Table 2 below shows the model results using all the genes given into the model. As shown in the table, there is a 100% AUC among random forest and KNN. This is indication that RNA sequence counts data clearly has predictive power in predicting preterm births. However, with such a small sample size more testing should be done.

Table 2. Model Result of RNA Sequence Count Data

	AUC in %
Stanford Study- Hierarchial Clustering	81%
Decision Tree	70%
Random Forest	100%
KNN	100%

6.2 RNA Concentration Data Set

The feature importance results for the RNA sequence count data set is shown in Figure 4 and Figure 7. The threshold of 0.5 times the mean is used for reasons stated earlier. As seen in the figures, the highest count is considered of high explanatory power in the model. The top seven genes considered to be of high importance are RPL23AP7, S10DA8, TBC1D15, OAZ1, POLE2, RAB27B, and DAPP1. The result here is similar to two genes from previous study where they found CLCN3, DAPP1, MAP3K7CL, MOB1B, PPBP, RAB27B, AND RGS18 as the main genes [1].One thing to note, the figure shows that the gene with

the highest count is 95,000 out of a 100,000 run. This further supports the instability in the feature importance when using Extra Trees classifier.

Fig. 7. Feature Importance of Genes in RNA Concentration Data for 100,000 Run

In Table 3 below, there is a wider range of accuracies compared to the RNA sequence count data set. However, the Decision Tree model shows the highest accuracy with 80% using 49 genes in the model. This result is close to Stanford's 81% AUC result. This is an indication that RNA concentration data may not have as much predictive power compared to RNA sequence count data. However, it still has merit for further investigation.

Table 3. Model Result of RNA Concentration Data

	Accuracy in %
Decision Tree	80%
Random Forest	77%
KNN	67%

7 Ethical Considerations

In data science, there are many ethical considerations to keep in mind in a study like this. According to Alan Fritzler [10], the Scope of the Project, Data Collection, Analysis, and Implementation are key elements to check for an ethical study. Fritzler addressed that the Project Selection and Scope of the project is important to note as we must evaluate the problem and determine if it is a symptom of a bigger issue.

In regards to the scope of this study, we are trying to predict spontaneous preterm birth from cfRNA. It is important to note that the cfRNA measurements intent is to only predict preterm birth, but the study fails to explain causality of why the spontaneous preterm birth occurred in the first place which is an important problem to solve. Though the study does not solve this issue directly, it may be the stepping stone to for understanding the circumstances of preterm birth.

Fritz [10] also mentioned that Data Collection is extremely important in regards to safeguarding privacy and having full disclosure of the subjects. All the women in all three cohorts were recruited[1], but we are unsure in regards to how much disclosure was given. Also, we have found by looking at the data set that there is no personable identifiable information. The data set contained race information for all three cohorts which may be influential to the study given that RNA is being measured. In regards to the study, there seems to be a lot of bias given that there are no preterm births from Caucasian or Hispanic women at all. Also, all preterm births came from only two cohorts with all African-American women which some had a history of preterm birth.

This race bias is concerning given that the population was small and that the population does not generalize to the rest of the population. In the Ngo, T., Moufarrej, M., et al. study, this was stated, "Our study has important limitations. Before a diagnostic or screening test based on this work can be used in the clinic, a blinded clinical trial with a larger sample size and diverse ethnicities is essential. Our pilot studies included one Caucasian cohort and two African-American cohorts; data from other ethnic groups would be valuable."

In the U.S. the race that delivers preterm the most often is African-American women according to the March of Dimes prematurity progress report. African American women deliver preterm 1.5 times more often than Hispanic and Caucasian women [11]. Some researchers tend to believe that Vitamin D may be one of the causes of preterm deliveries among African American women. Vitamin D is essential to the regulation of the immune system and insufficiency is linked to preterm birth. [12] However, several other factors also contribute such as diet, access to health care, socioeconomic status, microbiome, etc [11]. For the purposes of this study and the Ngo, T., Moufarrej, M., et al. study, no other considerations listed above were taken into account and there is no mention or data contributing to these other important factors other than progesterone was given to women showing signs of preterm labor.

8 Conclusion

When machine learning was implemented the AUC increased to 100% in comparison to the Stanford Study of 81%. This could be an indication that the gene list used may have predictive power in determining preterm births, especially in the case of RNA sequence count data. However, it should be noted that with such a small sample size over-fitting may occur even though measures have been taken to mitigate this issue. Therefore, there is a possibility that none of the genes have predictive power in differentiating preterm and full term births. Since the previous study only used seven features, the model estimation should contain at most seven features to see if the model estimation is more accurate.

In the case of the RNA concentration data set, there are slightly more observations than the RNA sequence count.Even though RNA concentration data set containing two cohorts was not as predictive as the RNA sequence count of one cohort. RNA concentration data may have potential for further study since one model was close to the Stanford result AUC. In both RNA sequence counts and RNA concentration data, A more diverse sample should be collected to confirm results. This allows not only for robust results, but applicability to the representative population.

References

1. Ngo, T., Moufarrej, M., et al.: Noninvasive blood tests for fetal development predict gestational age and preterm delivery. In: Science 2018, vol.360 pp.1133-1136. https://doi.org/10.1126/science.aar3819
2. Standford Medicine Newscenter, https://med.stanford.edu/news/all-news/2018/06/blood-test-for-pregnant-women-can-predict-premature-birth.html.Last accessed 5 Feb 2019
3. Ngo, T., Moufarrej, M., et al.: Supplementary Materials for Noninvasive blood tests for fetal development predict gestational age and preterm delivery. In: Science 2018, vol.360 pp.1133-1136. https://doi.org/10.1126/science.aar3819
4. Adi L. Tarca, Roberto Romero, Zhonghui Xu, Nardhy Gomez-Lopez, Offer Erez, Chaur-Dong Hsu, Sonia S. Hassan Vincent J. Carey: Targeted expression profiling by RNA-Seq improves detection of cellular dynamics during pregnancy and identifies a role for T cells in term parturition. In: Scientific Reportsvolume 9, Article number: 848 (2019). https://www.nature.com/articles/s41598-018-36649-w
5. ThermoFisher Scientific. https://www.thermofisher.com/us/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/basic-principles-rt-qpcr.html. Last accessed 23 Mar 2019.
6. Bustin, S. (2006). A-Z of quantitative PCR. La Jolla, CA: International University Line.
7. Human Genome Resources at NCBI. https://www.ncbi.nlm.nih.gov/genome/guide/human/. Last accessed 23 Mar 2019.
8. ThermoFisher Scientific. https://www.thermofisher.com/us/en/home/life-science/pcr/real-time-pcr/real-time-pcr-learning-center/real-time-pcr-basics/real-time-pcr-understanding-ct.html. Last accessed 23 Mar 2019.
9. Koh, W., Pan, W., et al.: Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. In: Proceedings of the National Academy of Sciences of the United States of America 2014 vol.111 .pp.7361-7366 https://doi.org/10.1073/pnas.1405528111
10. Fritzler, Alan. Data Science for Social Good. Center for Data Science and Public Policy, University of Chicago https://dssg.uchicago.edu/2015/09/18/an-ethical-checklist-for-data-science/
11. March of Dimes. Premature Birth annual Report Card: The March of Dimes is leading the Prematurity Campaign to reduce the nations preterm birth rate to 9.6 percent or less by 2020. US: 2014.
12. Sara A Mohamed,1,2,* Chandra Thota, Paul C Browne, Michael P Diamond, and Ayman Al-Hendy. Why is Preterm Birth Stubbornly Higher in African-Americans? https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402979/R3
13. Scikit-learn.org. (2019). sklearn.tree.ExtraTreeClassifier scikit-learn 0.21.2 documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.tree.ExtraTreeClassifier.html [Accessed 4 Jun. 2019].
14. Brownlee, J. (2019). Bagging and Random Forest Ensemble Algorithms for Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/ [Accessed 4 Jun. 2019].
15. Liberman, N. (2019). Decision Trees and Random Forests. [online] Towards Data Science. Available at: https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991 [Accessed 4 Jun. 2019].

16. Harrison, O. (2019). Machine Learning Basics with the K-Nearest Neighbors Algorithm. [online] Towards Data Science. Available at: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 [Accessed 4 Jun. 2019].
17. Seif, G. (2019). The 5 Clustering Algorithms Data Scientists Need to Know. [online] Towards Data Science. Available at: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 [Accessed 4 Jun. 2019].
18. "Preterm Birth." World Health Organization. February 19, 2018. Accessed July 07, 2019. https://www.who.int/news-room/fact-sheets/detail/preterm-birth
19. Ceballos, Frank. "An Intuitive Explanation of Random Forest and Extra Trees Classifiers." Medium. July 17, 2019. Accessed July 18, 2019. https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b.
20. "Sklearn.feature selection.SelectFromModel." Scikit. Accessed July 19, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.feature selection.SelectFromModel.html.
21. "Sklearn.model selection.StratifiedShuffleSplit." Scikit. Accessed July 19, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.model selection.StratifiedShuffleSplit.html.
22. "Sklearn.metrics.roc auc score." Scikit. Accessed July 19, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.rocaucscore. htmlsklearn.metrics.rocaucscore.
23. Narkhede, Sarang. "Understanding AUC - ROC Curve - Towards Data Science." Medium. May 26, 2019. Accessed July 19, 2019. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.
24. Drouin, Alexandre, Gal Letarte, Frdric Raymond, Mario Marchand, Jacques Corbeil, and Franois Laviolette. "Interpretable Genotype-to-phenotype Classifiers with Performance Guarantees." Nature News. March 11, 2019. Accessed July 20, 2019. https://www.nature.com/articles/s41598-019-40561-2.
25. Brooks, Eric L., and Ryan D. Kappedal. "Compressive Sampling for Phenotype Classification." 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017. doi:10.1109/bibm.2017.8217943.
26. Davis, James J., et al. "Antimicrobial resistance prediction in PATRIC and RAST." Scientific Reports 6 (2016): 27930.
27. Santerre, John W., et al. "Machine learning for antimicrobial resistance." arXiv preprint arXiv:1607.01224 (2016).Santerre, John William. Machine Learning for the Genotype-to-Phenotype Problem. 2018.