Predicting Premature Birth Risk with cfRNA

Jason Lin[1], Jonathan Marin[1], and John Santerre (Ph.D)[2]

[1] Southern Methodist University, Dallas TX 75205, USA
[2] University of Chicago, Chicago IL 60637, USA

Abstract. Determining important genes expressed in blood samples from pregnant women to provide an early indicator for the risk of preterm births is one of the elusive problems doctors have faced for many years. Previous methods were based solely on statistical tests and using this statistical result to see predictive error on the testing samples. Using cell-free RNA data from non-invasive blood tests and additional machine learning and statistical techniques, we improve upon the current methodology by seeing if other genes expressed have better predictive power. Hence, the model accuracy is improved by using a bigger feature space from 81% to 88% of the validation data set. These results help identify the important genes that could be used as early indicators of preterm births.

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

1 Introduction

The ability to give doctors earlier indicators concerning the development of an infant can help doctors plan early treatment plans to help infants cross over to later weeks of viability with a better quality of life. Earlier weeks of delivery may incur complications to the infant and may affect viability. Even though countless studies and analysis have been done on the biology of fetal development, there has not been a proposed test that is both accurate and easy to implement to determine preterm birth. Present medical methods, mainly ultrasounds and the last menstrual period, are imprecise and are easily miscalculated and misinterpreted [1]. Also, these medical methods can only measure the gestational age of the baby and not the risks of being a preterm birth.

Ngo, T. and Moufarrej, M. et al. utilized cell-free RNA (cfRNA) data from pregnant women and proposed two models. The first model was a random forest regression model that utilizes 8 cfRNA to predict gestational age with high accuracy. This dataset only contained a Denmark cohort of Caucasian women who all delivered full-term. Ngo, T. and Moufarrej, M. et al. then attempted to use another cohort containg 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 high

accuracy[1]. However, some drawbacks of these models are that further tests need to be done on a larger population to ensure its accuracy and ensure a more robust results since the number of observations used is 215 for the gestational age study and 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.

The hierarchical cluster model analysis is the main problem of interest given that it tries to determine what genes are considered indicators for preterm birth risk. The hierarchical model deemed to have an accuracy 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. Therefore, we will explain this problem further by showing that other genes are better indicators for preterm risk and other methodologies yield better results.

We developed higher accuracy in two ways. First, we are creating lower dimensions for the gene set. Second, we are using a different model algorithm implementation. 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 (UPenn) and University of Alabama at Birmingham cohort (UAB) [1], and whether the use of principal components analysis (PCA), interaction, and feature engineering can increase the prediction power of the model with the rest of the cfRNA features. We dealt with the small sample issues by trying bootstrapping and other models that utilize Bayesian. Other techniques such as PCA, interaction, and feature engineering can decrease the dimensions of the data set and help explain variability which help combat overfitting especially in the case of such little data.

There are many different data mining, machine learning, and statistical techniques that can be implemented to increase accuracy 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. Other methodologies introduced were XGBoost, Clustering, and Ensembling to see if these models can produce higher accuracy. Feature importance analysis was also used to deal with what data should be included in the training and testing set.

Analysis of the RNA sequence count data using only the 15 women from the UPenn 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 accuracy of the models shows to be 100% when all features are used. This could be indication of overfitting and therefore the number of features should be cut down, or bootstrap or bagging should be used to mitigate this issue. Therefore at present no further models are attempted until this issue is resolved. However, apart from the overfitting issue, when looking at the top seven genes

considered to be of high importance in the feature importance methodology, it shows LYPLAL1, LACTB2 RRP1B, PCTY1B, MAL, SH3GLB2, and OSBPL5. 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].

Analysis of the RNA concentration using the 38 women from UAB and UPenn showed completely different results compared to the RNA sequence count data. The accuracy of the model when using decision tree is 0.75, random forest is around 0.625, logistic regression is 0.625, and K nearest neighbors (KNN) is 0.88 when using all 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 0AZ2, dCD, CD24, GAPDH, TBC1D15, RAB11A, and PGLYRP1. Which as stated earlier is completely different from the previous result as shown in Ngo, T. and Moufarrej, M. et al [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 when all genes are used in the model between the two data sets. However, this result is confounded by the small dataset 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.

2 Preterm Birth

The Ngo and Moufarrej et. al. study defines preterm birth as the following: delivery at ¡37 completed weeks of pregnancy. Spontaneous preterm birth includes preterm labor, preterm spontaneous rupture of membranes, preterm premature rupture of membranes (PPROM) and cervical weakness [1].

3 Summary of Hierarchical Study

The Ngo and Moufarrej et. al. study investigated what 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 not [1]. The methodology used to determine the 38 potential genes is using 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 of 3 genes for each of the 13 combinations used this to validate the UAB and Denmark cohorts with an area under the curve of 0.81 and 0.86 respectively.

4 Data Collection

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

The University of Alabama at Birmingham 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]. University of Pennsylvania contains 15 women that were studied where one had preeclampsia [3]. In the RNA sequence count data, only the University of Pennsylvania cohort was used in the hierarchical study [1,3]. The RNA concentration data set, used as another potential data set which is not used by Ngo and Moufarrej, is the combination of University of Alabama and University of Pennsylvania, where in University of Pennsylvania two women were dropped from the group [3].

This information is important when it comes to the interpretation of the results from the study. These exogenous variables may have influence in the results since medical history of the women may influence the probability of the women having the same issue in the future. Therefore, variables such as this should be kept in mind since this can have interactions with other variables influencing what genes are considered important in the model.

The main instrument used to determine what genes and their given levels that may have predictive power in determining preterm births is blood samples. In both the University of Alabama at Birmingham and University of Pennsylvania 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 method 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]. 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. 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]. In the study Koh and Pan et. al., it 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, further investigation should be done to see if other genes have importance.

In the data set for concentration of RNA expression, it contains the 23 women from the University of Alabama at Birmingham and the 13 women from the University of Pennsylvania. The methodology described above is what 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 University of Pennsylvania 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]. With these blood samples researchers can conduct a liquid biopsy in order to predict preterm delivery. RNA transcript profiling using micro-arrays and RNA sequencing takes measurements of thousands of protein-coding and non-coding genes[4]. RNA samples were collected from two cohorts: the University of Pennsylvania, and the University of Alabama at Birmingham. [1]

Below are the summary statistics of the two cohorts used. The University of Pennsylvania consisted of 15 women that were at risk of preterm birth because of symptoms shown, however 7 women delivered at full term and the other 8 at preterm. Samples for the Pennsylvania cohort were only collected once and at the time of delivery. The cohort for the University of Alabama at Birmingham had 26 pregnant women in which only five had delivered preterm spontaneously and eighteen have delivered to full term. [1] The women from both Alabama and Pennsylvania 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) preterm	Alabama (n=18) - full-term	Alabama (n=5) - preterm
Age(years, mean)	23.7	23.0	25.28	25.8
BMI(kg/m2, mean)	31.9	25.1	28.6	33
Ethnicity - Hispanic(%)	0	0	0	0
Ethnicity - Caucasian(%)	0	0	0	0
Ethnicity - African-American(%)	7	8	17	5
Gestational Age at Delivery (weeks, mean)	39.4	26.4	38.7	30.6

5 Feature and Sample Methodology

5.1 Feature Engineering, Feature Importance, and Dimension Reduction

Using the same two data set provided by Ngo and Moufarrej et. al. study, feature engineering has been attempted on this data set. The 3 gene group for 13 combinations is one feature, that is proposed by the previous study, to improve accuracy in the model building such as random forest and logistic. From a statistical point of view, the gene expressions alone may not have explanatory power, but as a group may have significant explanatory power. Therefore, this grouping proposed in the previous paper may have statistical significance. Another feature engineering is to look at different forms of the features by squaring, cubing, raising to the power of the fourth etc. By creating and taking account of these different functional forms, the model can take account for non-linear relationship that may not be found at initial estimation.

Principal Component Analysis is also used to create new features and reduce the feature dimension for the data set. In PCA, the algorithm essentially tries to create a linear combination of the features called principal components, where the number of principal components is defined by the user. The assumptions of this algorithm are that each principal component is uncorrelated with one another and that the correlation in the given principal component is maximized [13]. Being that there is such a small data set, a PCA ranging from 3 to 7 is used to see if there is any significant increase in accuracy when grouped by this algorithm.

In concern to the creation of these many features, feature importance can help determine what features have better explanatory power than another. Using scikit learns Extra Tree Classifier, features are 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 to select the top features to use in order to limit the risk of overfitting and runtime.

5.2 Training/Test Splits, Bootstrapping, and Bagging

In the original study, the training data set was one entire cohort (Pennsylvania) while the testing set was a different cohort (University of Alabama Birmingham). The same train and test split methodology for model fit comparison will be used for the RNA sequence count data set. However, for the RNA gene expression concentration, the two cohorts are combined, and a stratified random sample based on the response (preterm/full-term) is used for the train test split.

In the both data sets, bootstrapping is used to mitigate the small data issue. Bootstrapping is essentially a re-sampling technique that allows for the inference of certain population statistics [15]. The bootstrapping methodology is essentially a random draw with replacement to a certain sample size from a given pool of data designated by the user. This allows for the creation of a

distribution to allow for inference on the population statistics. The main way bootstrapping is used in machine learning and modeling is to estimate multiple models on the different bootstrapped samples and then taking an average of all the results to see what type of prediction can be seen. Bagging is essentially the same concept as bootstrapping but is mainly used in CART models as a way of ensembling so that one tree does not over influence the results [16]. Bagging is essentially running the decision trees on the bootstrapped training data, and then using an aggregation methodology to determine the final decision of the model, which can be the average, most frequent vote, etc.

6 Classification Algorithms

Once the data is cleaned and the features are created, the following section are the background information on the algorithms and methods used to determine the gene 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.

6.1 Logistic Regression

The logistic regression is a statistical model that provides a probability for a given event to happen. The response variable form that the logistic regression typically uses is in the forms of zeros and ones. Therefore, when this response variable is plotted with the given features, it follows a form that is close to a sigmoid i.e. logit function as seen in Figure 1 below.

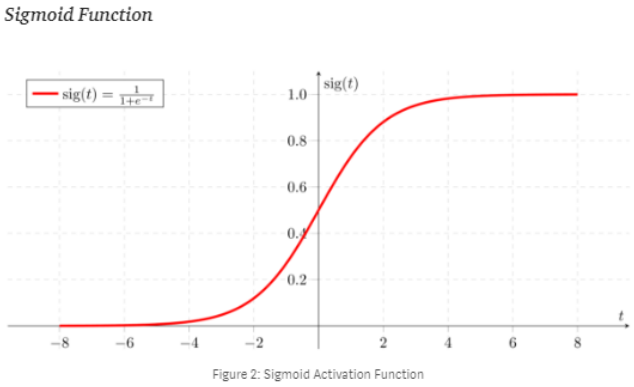

Fig. 1. Logistic Function [17]

From here the function can be converted to a linear function by applying the logit formula shown in the figure. Therefore, the link between the features

and the given classification is explained by this linear function. Using this form is more suitable than the linear regression, where in the linear regression model the range is not bounded, and the slopes are not easy to interpret [17].

6.2 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 [18]. 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. 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 [18]. 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 [18].

Therefore, random forest is used to combat these issues that the decision tree has by creating multiple decision trees. By creating multiple decision trees, random forest is able to reduce variance by training on different samples of the data [18], which means that the variability in the optimal result because of reduced depth is mitigated since this can be re-estimated many times through random forest. Also, with random forest, the model 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 [18]. 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.

6.3 AdaBoost and Gradient Boosting

Boosting methods are a way to take certain models that have poor performance and re-estimate the model through an iterative process in order to increase accuracy and performance. There are two methods that are used in this paper to see whether these methods can help find gene indicators that may have been overlooked because of this weak learner problem [19].

The first boosting method is AdaBoost, or also called Adaptive Boosting. The method is simple in that AdaBoost takes weak learners, which are decision trees with a single split, called decision stumps for their shortness. [19] and weights these observations to the group that is difficult to classify. As more weak learners are made, they are subsequently added to the group and weighted to train on the difficult class. The predictions made by these weak learners are by majority vote where the vote is weighted on each weak learners accuracy for the result [19].

The next method is Gradient Boosting, where the main difference is that gradient descent is used to optimize the model. In Gradient boosting, the loss function must be specified, where the form is differentiable such as the squared error in multiple linear regression [19]. This means that the loss function is used to determine the parameters of the given model by minimizing function such a minimizing error in linear regression etc. After the loss function is specified, the weak learner is then specified where in Gradient Boosting it does not have to be single split but can be somewhat larger tree [19]. Then the weak learner trees are added to the model, and the gradient descent algorithm is used to minimize the loss function by taking a step in a direction towards the minimum where then the trees are then re-parameterized.

6.4 Clustering Methodologies

Clustering methods are pursued because of the ease of implementation, but also there is no data size requirement to run this methodology. 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 how where the points cluster for preterm and full-term births. The main cluster methods used in this paper are: K Nearest Neighbors (KNN), K-Means, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

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 [20]. The main drawback of this method is that computational time may increase as the data set becomes larger.

In K-means, it is like KNN, however, the main difference is that number of classes are first initialized. This initialization determines how many random points should be placed in the data field. From these random points, an iterative process then occurs to find the center of data cluster starting from the random points position. Then after the iterations are completed, the groups are then specified based on where the center points of the data are found [21].

In DBSCAN, this process is a computationally intensive process in that each iterative process updates what is considered a cluster. Here a random point is placed and then distance metric is calculated called epsilon. If there are enough data points around the random point as specified by the minPoints and within epsilon, then this cluster is considered a cluster otherwise it is considered noise [21]. The points within that cluster is then determined to see if other points are within epsilon and are then considered part of the cluster. This iterative process is done until all point within epsilon are taken accounted for. Then a new random point is places and goes through the same process until all the data points have gone through the algorithm [21].

7 Results

7.1 RNA Sequence Count Data Set

The feature importance results for the RNA sequence count data set is shown in Figure 2 below. As seen in the figure, the highest Gini Importance measure is considered of high explanatory power in the model. The top seven genes considered to be of high importance are LYPLAL1, LACTB2 RRP1B, PCTY1B, MAL, SH3GLB2, and OSBPL5. 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].

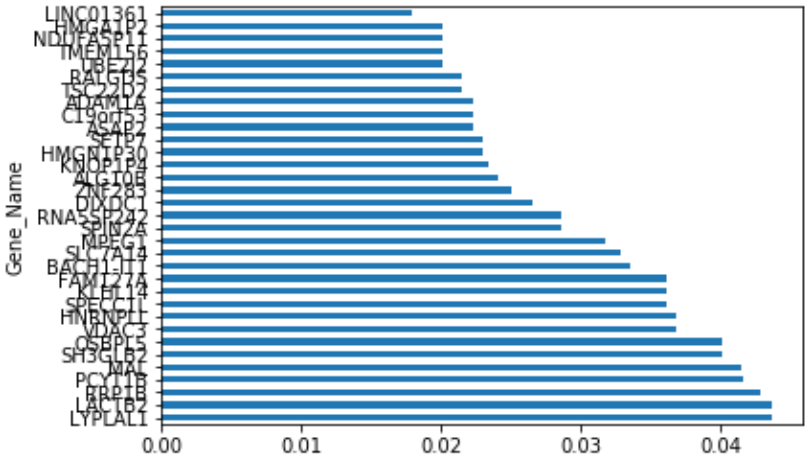

Fig. 2. Feature Importance of Genes in RNA Sequence Count Data

In Figure 3 below, shows the decision tree used in the RNA Sequence Count data, showing the depth of the tree. It shows the gene PPBP is used as the main driver for differentiating between preterm and full-term birth.

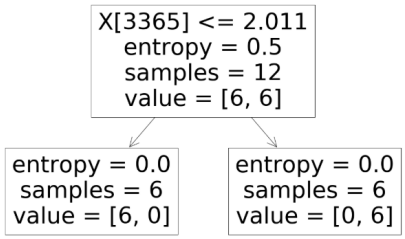

Fig. 3. Decision Tree of Gene in RNA Sequence Count Data

The Table 1 below shows the model results using all the genes given into the model. As shown in the table, there is a 100% accuracy among all the different methodologies/

Table 1. Model Result of RNA Sequence Count Data

	Accuracy in %
Decision Tree	100%
Random Forest	100%
KNN	100%

7.2 RNA Concentration Data Set

The feature importance results for the RNA sequence count data set is shown in Figure 4 below. As seen in the figure, the highest Gini Importance measure is considered of high explanatory power in the model. The top seven genes considered to be of high importance are 0AZ2, dCD, CD24, GAPDH, TBC1D15, RAB11A, and PGLYRP1. 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].

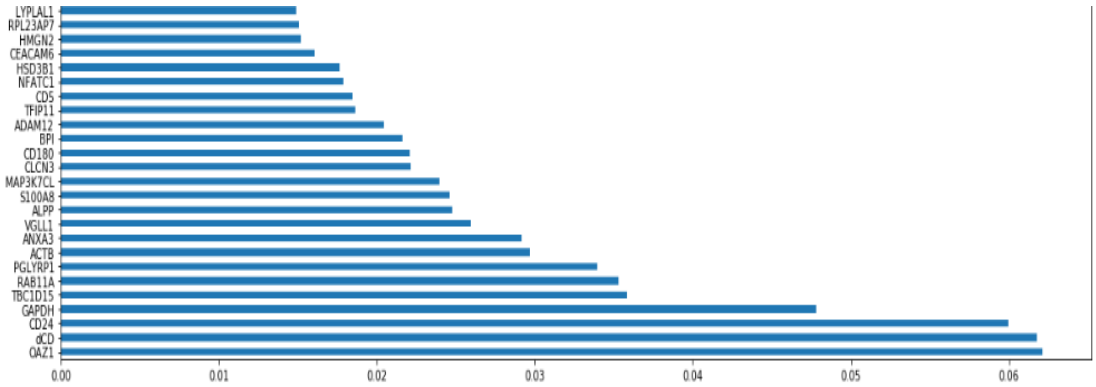

Fig. 4. Feature Importance of Genes in RNA Concentration Data

In Figure 5 below, shows the decision tree used in the RNA Concentration data, showing the depth of the tree. It shows there are 4 genes that are the main driver for differentiating between preterm and full-term birth.

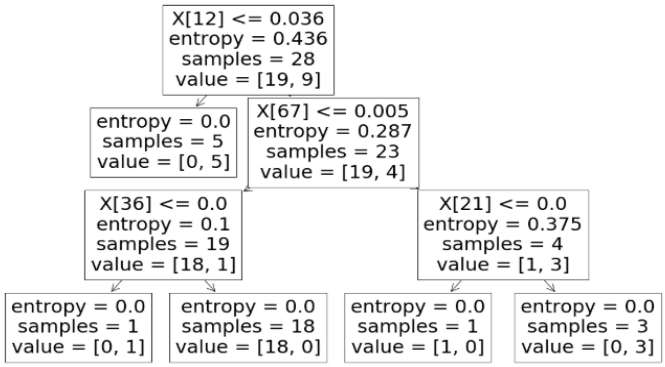

Fig. 5. Decision Tree of Gene in RNA Concentration Data

In Table 2 below, there is a wider range of accuracies compared to the RNA sequence count data set. However, the KNN model shows the highest accuracy with 87.5% using all the genes in the model.

Table 2. Model Result of RNA Concentration Data

	Accuracy in %
Decision Tree	75%
Random Forest	62.5%
AdaBoost	50%
Gradient Boost	50%
Logistic	62.5%
KNN	87.5%

8 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.

9 Conclusion

As seen in the results, the RNA sequence count data set seems to show some issues with low observation count. As shown seen in the accuracies of the models estimated, all of them show a 100% accuracy. This seems to indicate that overfitting maybe occurring in the data, and therefore bagging and bootstrapping should be used to obtain a more accurate result or less features should be used.Therefore, there is high possibility that none of the genes have preditive 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. Therefore, the accuracies are wider

in range, where it seems to show that KNN has the highest accuracy when all genes are used in the model. This is indication that at least one of the genes has predictive power in differentiating between preterm and full-term. Therefore, the analysis should look at it from a gene by gene perspective.

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. Lever, J., Krzywinski, M. and Altman, N. (2017). Points of Significance: Principal component analysis. Nature Methods, 14(7), pp.641-642.

14. 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].
15. Brownlee, J. (2019). A Gentle Introduction to the Bootstrap Method. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/ [Accessed 4 Jun. 2019].
16. 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].
17. Swaminathan, S. (2019). Logistic Regression Detailed Overview. [online] Towards Data Science. Available at: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc [Accessed 4 Jun. 2019].
18. 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].
19. Brownlee, J. (2019). A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/ [Accessed 4 Jun. 2019].
20. 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].
21. 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].

A Appendix

Still need to do analysis on bootstrapping and bagging.

Also use fewer features in the model estimation

Need to do boosting and cluster methods on the other data sets.