Application of Dimensionality Reduction (DR) techniques in the diagnosis of Fetal Health

Swara Lawande
CIS6930
Gainesville, FL, USA
lawande.s@ufl.edu

Delora Almeida
CIS6930
Gainesville, FL, USA
deloraalmeida@ufl.edu

Ashutosh Yadav
CIS6930
Gainesville, FL, USA
ashutosh.yadav@ufl.edu

Naman Bhatia
CIS6930
Gainesville, FL, USA
naman.bhatia@ufl.edu

***Abstract*—"Fetal mortality is an important public health issue that is sometimes neglected." According to the "National Center for Health Statistics", there are around 1 million fetal deaths in the United States each year. Even though the mortality has been decreasing for years, the fetal death rate in the United States has saturated at a high number. Fetal health has a direct impact on maternal health as well. The introduction of use of "cardiotocograms" has made critical data collection easier yet correctly diagnosing conditions is tedious. A combination of digital health and machine learning technologies are being used worldwide to detect and prevent a lot of diseases. Using our knowledge of various existing data science techniques and the data set provided to us that contained real life CTG exams along with their fetal health classifications by experts, we have developed an efficient algorithm to make automated assessment of fetal health conditions possible.**

***Index Terms*—Data Mining, Disease Diagnosis, Classification, Health Data, Data Science, Feature Selection, Principal Component Analysis, Dimensionality Reduction**

I. Introduction

Data in health care is a valuable resource that can be used to predict requirements in the future of health and conducting clinical research. However, the complexity and large amounts of medical data limit applications of data mining techniques. In this project we have developed a methodology that reduces the dimensionality of a fetal health dataset.

A fetus is basically an unborn baby from eighth week of pregnancy till it is born. Body defects and fetal illnesses are examples of fetal conditions, which are a range of issues that impair the health of an unborn infant. A lot of research is being done to improve the baby's health and well-being before as well as after birth. With the advancement in digital health technologies, researchers and health care workers have been able to collect data easily on the fly but flooding the decision makers with a lot of data is not the end of the solution. Making sure that capturing only the information that adds value in the decision making/diagnosis is important as it will save time and resources, but not at the cost of accuracy since we're risking lives here.

"Cardiotocography is a simple and inexpensive way for healthcare providers to examine fetal health and take measures to reduce infant and maternal mortality". The machine sends ultrasound pulses and reads the response, providing real time information on the fetal heart rate, decelerations, movements, uterine contractions and many other metrics. Now since we have a lot different sets of values, using prediction algorithms to diagnose becomes trickier since high dimensionality comes with its own set of problems like high computation costs, over fitting and most importantly low accuracy.

Dimensionality reduction refers to techniques that reduce the number of features in a dataset. In machine learning, a dataset may contain multiple features which makes it more difficult to model them, also known as the "curse of dimensionality". A machine learning model trained on a large number of features, overfits by getting dependent on the data it was trained on, which results in poor performance on real input data. Lesser the features the training data has, the simpler the model will be. Dimensionality reduction has more benefits such as it removes noise and redundant features, thereby improving model performance. Lesser the amount of misleading data, better the model accuracy. Fewer dimensions leads to lesser time for computations and also data compression, so storage space required is decreased.

For dimensionality reduction, we have implemented five classification algorithms along with feature selection and extraction methods. The papers we reviewed, they used mainstream feature selection methods and the classification algorithms. As mentioned, with the implementation of these five algorithms we can compare the results obtained, and determine which one provides the best results for our dataset based on fetal health. We have also employed the most renowned extraction method Principal Component Analysis (PCA). Our results show that most of the classification algorithms provide better accuracy with the feature selection, followed by feature extraction implementation of PCA.

In this paper, we have implemented classification algorithms with feature selection and feature extraction which are two of the main types of dimensionality reduction techniques.

II. Motivations

United Nations has mentioned about reduction of child mortality in its sustainable goals and consider it as a key indicator of human progress. Maternal mortality is directly related to the concept of fetal health and is responsible for 295,000 fatalities during and after pregnancy. CTG exams have provided almost all the critical features to diagnose fetal health but having a lot of information sometimes makes automation a lot more difficult.

Adding new features to the data-set can help improving the machine learning model, but too many input variables can dramatically degrade the performance of machine learning algorithms. The computational cost of performing learning can be high when there is a high dimensionality. A model will frequently over-fit when being learned, meaning that it will do well when learning it but perform poorly when testing it. There are very few random data distributions in high dimensions and a high degree of correlation, often with spurious correlations. Several distance-based analysis tools are susceptible to the problem of equivalence of distances in high dimensions, which can affect their accuracy.

III. Case of study

Fetal health conditions refer to problems that affect an unborn baby's health. It may be birth defects or other fetal disorders. This can be monitored for better action in case of an unforeseen problem. Fetal health monitoring helps display any risks to the fetus during late pregnancy, high risk pregnancy or labor, for instance if a C-section might be needed. It is done to maintain the health and well being of the baby. Cardiotocography (CTG) is a method used to monitor the fetal-maternal health data such as the fetus's heartbeat and the uterine contractions during pregnancy. There is a very low chance of physical risk with this method and a lot of data related to the fetus's health is obtained.

It is a necessity to reduce false negatives in predictions. For instance, misclassification of a member of the 'Normal' class as 'At Risk' is a small issue, however a way more serious consequence is misclassifying a member of the 'At Risk' class as 'Normal'. Our goal is to ensure that we find fetus's who are in trouble so that decisions can be made accordingly.

The dataset that we worked with consists of 2126 feature records of 29 features collected from various cardiotocogram tests [1], which were then grouped into three categories ("1-Normal", "2-Suspect" and "3-Pathological") by various expert obstetricians.

The features of the fetal health dataset are as mentioned below:

baseline value	Baseline Fetal Heart Rate (FHR)
blood_glucose	Bood glucosa levels (mg/dl)
bpm	Beats per minute
accelerations	Number of accelerations per second
accelerations_per_minute	Number of accelerations per minute
fetal_movement	Number of fetal movements per second
fetal_repose	Percentage of time of fetal repose per second
uterine_contractions	Number of uterine contractions per second
percentage of uterine contractions	Percentage of uterine contractions per second
light_decelerations	Number of LDs per second
severe_decelerations	Number of SDs per second
prolongued_decelerations	Number of PDs per second
medium_decelerations	Number of MDs per second
mean_decelerations	Decelerations mean
abnormal short term variability	Percentage of time with abnormal short term variability
mean value of short term variability	Mean value of short term variability
Percentage of time with abnormal long term variability	Percentage of time with abnormal long term variability
percentage of time without abnormal long term variability	Percentage of time without abnormal long term variability
mean value of long term variability	Mean value of long term variability
histogram width	Width of the histogram made using all values from a record
histogram min	Histogram minimum value
histogram max	Histogram maximum value
histogram number of peaks	Number of peaks in the exam histogram
histogram number of zeroes	Number of zeroes in the exam histogram
histogram_mode	Hist mode
histogram_mean	Hist mean
histogram_median	Hist median
histogram_variance	Hist variance
histogram_tendency	Histogram trend
fetal_health	Normal, Suspect, Pathological

Table 1. Feature description

IV. Proposal of solution

The below figure represents the flow of data mining and data science methods applied in our project.

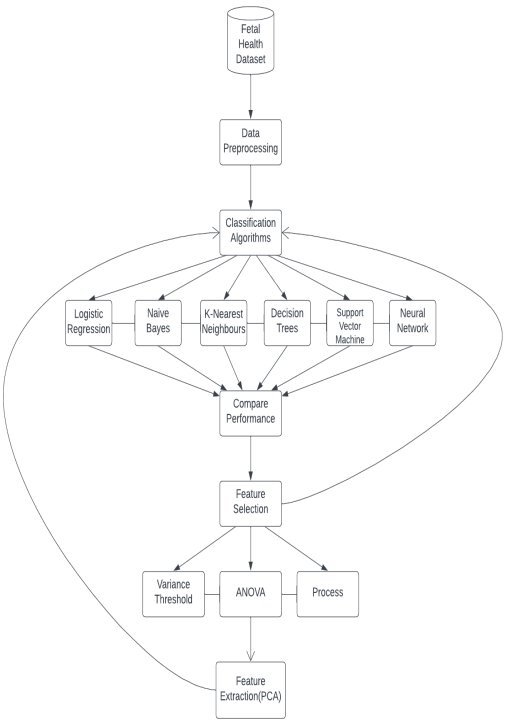

Fig 1. Process Flow

Our project's main goal is to perform dimensionality reduction on the given data-set. We have used a combination of techniques such as feature selection and feature extraction before applying the various machine-learning models, so as to accurately evaluate fetal health risks.

- **Data Pre-processing:**
  - Firstly, the dataset was separated into feature set - X and the label - Y ('fetal_health').
  - The data was filtered for any missing or invalid values.
  - Further, we used the Standard Scaler from sklearn library to normalize the features so as to bring all the features within a common range. This step is crucial for classification, specially before performing feature extraction techniques.
  - Finally, we split the data into training and testing sets with 20% test data.
- **Feature Selection:**
  This step helps in removing the less relevant parameters from the feature set. We have tested the following feature selection techniques in our project:
  1) Variance Threshold: Variance indicates how spread out the data is from the mean. Features with less variance offer less information, hence we can remove such features while performing classification. We use scikit-learn's VarianceThreshold for this step.
  2) ANOVA: 'Analysis of Variance' is a parametric hypothesis test for determining whether the means from two or more samples of data come from the same distribution or not. [2] ANOVA is effective specially when the features are numerical and the target label is categorical, which is the case with our dataset.
  3) Pearson's Correlation Coefficient: In this method, the association between the individual features and the target label is measured using the correlation coefficient metric. It is based on the method of covariance and provides a good statistic for eliminating features based on a correlation threshold.
- **Feature Extraction:**
  This is a crucial step in Dimensionality reduction, wherein new features are extracted from the existing ones. This new feature set must represent the existing feature set in such a way that there is negligible loss of information, that is, most of the information in the original data set must be covered.

  In our project, we have used Principal Component Analysis (PCA) for Feature Extraction. PCA aims to maximize the feature variance and minimize the reconstruction error. We have used scikit-learns PCA library to implement this step.

- **Classification:**
  We have used the following machine learning algorithms to perform classification:
  1) Logistic Regression: This algorithm uses the sigmoid function to perform classification. It is more popular in binary classification problems.
  2) Naive Bayes: The Naive Bayes classifier is based on the assumption that the features are independent from each other, that is, they are not correlated. This is a well-known multi-class classifier.
  3) K-Nearest Neighbours: This algorithm is based on the assumption that similar things reside close to each other. It performs proximity analysis of the different feature sets to perform classification.
  4) Decision Tree: A decision tree is built using binary recursive partitioning. The location of attributes in the tree are decided using the Gini index, which is a measure of impurity in the attribute. An attribute with low Gini index is preferred over an attribute with high Gini index.
  5) Support Vector Machine: For a multi-class classification problem with n classes and m attributes, this algorithm creates m-dimensional hyperplanes dividing the space into m partitions. Support vectors are the data points that are close to the hyperplanes.
- **Performance Evaluation:** We have used 2 metrics for evaluating the performance of the aforementioned machine learning models: Accuracy and F1-score.

  The following terms are required to compute these metrics:

- TP : True Positive
- TN : True Negative
- FP : False Positive
- FN : False Negative

The formula for computing are shown in fig.2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Fig 2. Performance Evaluation Metrics

Accuracy represents the ratio of correctly classified instances, while the f-1 score gives a better measure of the incorrectly classified cases. Hence, we use both the metrics for evaluating the models.

V. Results

- **Initial Results**
  We trained the machine learning models using the preprocessed data without applying any dimensionality reduction methods. The performance of the models on test data is reflected from the below table:

Algorithm	Accuracy	F1-Score
Logistic Regression	0.8808	0.88058
Naive Bayes	0.7272	0.7577
K Nearest Neighbours	0.8855	0.8844
Decision Tree	0.9310	0.9313
Support Vector Machine	0.9090	0.9086

Table 2.1. Accuracy in Initial Results

- **Feature Selection**
  For feature selection, we selected three methods as described in section IV.

  The below image shows the features with a correlation coefficient more than 0.2. It is evident that certain attributes such as 'prolonged decelerations' and 'abnormal short term variability' are strongly correlated with the label and thus must be retained.

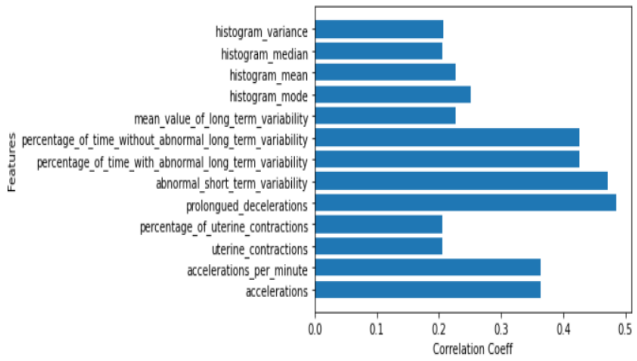

Fig 3. Correlation of features with respect to fetal_health

Further, we compared the effect of each method on the model performance.

- Threshold Variance led to the elimination of 10 features (19 features retained).
- In ANOVA, we observed that using k=20 let to the best results, that is 9 features were eliminated.
- In Correlation Coefficient, we chose 0.25 as the threshold for elimination of features. 7 features were retained in the end.

Clearly, ANOVA is the best suited for our application and this is evident in the below figure 4. It is also a popular feature selection method when the classifying attributes are numerical and the target labels are categorical.

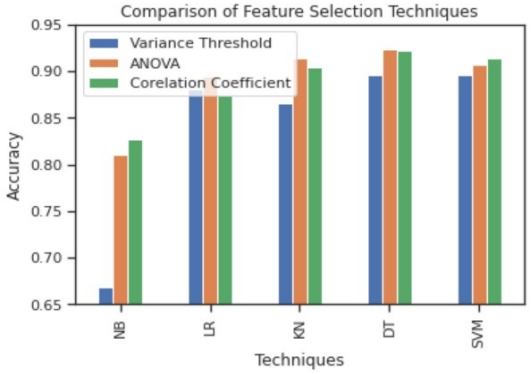

Fig 4. Comparison of feature selection techniques

The accuracy and f1-score of the ML models after applying ANOVA are shown in the below table.

Algorithm	Accuracy	F1-Score
Logistic Regression	0.8934	0.8946
Naive Bayes	0.8056	0.8233
K Nearest Neighbours	0.9043	0.9038
Decision Tree	0.9278	0.9266
Support Vector Machine	0.9184	0.9174

Table 3.1: Accuracy after Feature Selection

- **Feature Extraction** We used scikit-learn's PCA library to further perform Principal Component Analysis on the feature set obtained after Feature Selection step. We observe that this step does not impact the accuracy of the models, but the f1-score is improved. This indicates

that PCA reduces the model's tendency of incorrect classification.

Algorithm	Accuracy	F1-Score
Logistic Regression	0.8934	0.8946
Naive Bayes	0.8369	0.8482
K Nearest Neighbours	0.9043	0.9038
Decision Tree	0.9278	0.9266
Support Vector Machine	0.9184	0.9174

Table 4.1. Accuracy after Feature Extraction

- **Performance Comparison**
  The below image shows the affect of applying the different dimensionality reduction techniques to the dataset. The accuracy of the ML models is plotted for every model in Fig 5.

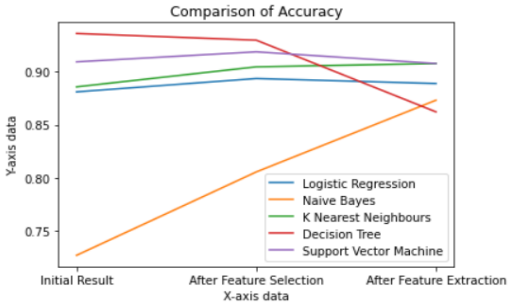

Fig 5. Comparison of ML Algorithms

We observe that the performance of all the models improves greatly after dimensionality reduction, except for Decision Tree. For the rest, the performance increases greatly after applying feature selection (ANOVA) and slightly after feature extraction (PCA).

VI. Discussion

We observe that it is possible to reduce the number of features, i.e., perform dimensionality reduction on a fetal health dataset. We employ the classification methods Logistic Regression, Naive Bayes, KNN, Decision Tree and Support Vector Machine. This was followed by feature selection with classification algorithms. With respect to fig.4, we observe that the best accuracy is achieved with Decision Tree under Anova technique. We then follow this with PCA for feature extraction.

Our aim is to reduce false negatives. For instance, misclassifying a member of the 'Normal' class as 'At Risk' is a small issue, however a way more serious consequence is misclassifying a member of the 'At Risk' class as 'Normal'. Our goal is to ensure that we find fetus's who are in distress so that medical-related decisions can be made accordingly. Thus the models Recall/Sensitivity are prioritized to evaluate the models. This will ensure that we capture the positive class, thereby, we can avoid false negatives. We tested different algorithms for feature selection in this paper. Our intent wasn't to reduce the features to the bare minimum, but rather to eliminate those that don't have a significant impact on the output, so we can save on computing power while retaining high accuracy.

VII. Related Work

As prediction of a disease can help clinicians make more accurate decisions regarding a patients' health, machine learning can be used to understand the symptoms related to the disease.

Different algorithms can be used for reducing the dimensionality based on the feature selection methods. Zebari et al. [3] observed that the most used classification algorithms for feature selection were KNN and SVM with SVM achieving the best accuracy. In terms of feature extraction, techniques such as DNN and CNN can be used but PCA is used most widely as it achieves better accuracy, and reduces the number of features, thereby improving performance.

Datasets related to medical data contain data which has noise. Feature selection is used to remove irrelevant and redundant features [4]. Escamila et al. [5] proposed dimensionality reduction method by applying a feature selection technique to find features for heart diseases. They improved the prediction model of whether a patient has heart disease using PCA with chi-square which had an accuracy of 98%. Similarly, [6] presents a approach for hybrid feature selection to reduce dimesionality for a chronic disease dataset. The feature selection process generates a weight matrix and a list of features with weights below the specified threshold value. This is then passed to the PCA which assigns an f-score to every feature. For efficient chronic disease diagnosis [7], classification techniques such as SVM, decision tree, naive bayes, and random forest are used. Our aim is similar and we have performed dimensionality reduction on fetal health data with feature selection and extraction.

VIII. Conclusions and Future Research Lines

We have implemented the various classification, feature selection and feature extraction algorithms known to us and observed the difference in the performance of each. In this process, we achieved dimensionality reduction which helps us to train our models in a way so that the accuracy is improved while retaining all the critical information. We saw from our results that "prolonged decelerations" and "heart rate variability" are the most correlated features in the data given to the fetal health conditions. We came up with the our solution using the algorithms known to us, but if we use more advanced algorithms, we could achieve even better results. We also learnt that accuracy and f1 scores both together give us a better overview of the performance. Using our algorithm along with the cardiotocograms data would help healthcare workers to diagnose a lot of fetal health conditions easily.

Even though we achieved decent results with the given data and our algorithm, we feel if CTG is highly effective and

should be used more widely. When more data is fed into our algorithm, we can achieve even better results and never miss any instance of at risk fetal health. We recommend the use of CTG's more frequently to all pregnant women to preserve both maternal and fetal health.

REFERENCES

[1] D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de Sa, and L. Pereira-Leite, "Sisporto 2.0: a program for automated analysis of cardiotocograms," *Journal of Maternal-Fetal Medicine*, vol. 9, no. 5, pp. 311–318, 2000.

[2] J. Brownlee, "How to perform feature selection with numerical input data," Aug 2020. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-numerical-input-data/

[3] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, 2020.

[4] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 Science and Information Conference*, 2014, pp. 372–378.

[5] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and pca," *Informatics in Medicine Unlocked*, vol. 19, p. 100330, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820300125

[6] D. Jain and V. Singh, "An efficient hybrid feature selection model for dimensionality reduction," *Procedia Computer Science*, vol. 132, pp. 333–341, 2018.

[7] A. M. Alhassan and W. M. N. Wan Zainon, "Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis," *IEEE Access*, vol. 9, pp. 87 310–87 317, 2021.