Fetal Distress Classification Using Deep Learning Based on Cardiotocography Signals

Akshay Menon
Big Data Analytics Laboratory
Presidency University
Bangalore, India
Email: akshay.20201csd0033@presidencyuniversity.in

Harishkumar K S
Big Data Analytics Laboratory
Presidency University
Bangalore, India
Email: harishkumar@presidencyuniversity.in

Kushal V
Big Data Analytics Laboratory
Presidency University
Bangalore, India
Email: kushal.20201csd0057@presidencyuniversity.in

Harini S
Big Data Analytics Laboratory
Presidency University
Bangalore, India
Email: harini.20201csd0051@presidencyuniversity.in

***Abstract*—This paper focuses on developing a deep learning-based solution for the classification of intrapartum fetal distress using cardiotocography (CTG) signal data. Fetal distress diagnosis has traditionally been quite subjective, and our study aims to develop an intuitive and objective solution using deep neural networks, particularly, a one-dimensional convolutional neural network (CNN) model trained on two time series signals: Fetal Heart Rate and Uterine Contractions. We focus on identifying key combinations of neonatal criteria like umbilical arterial *pH* and *Apgar* scores, and their optimal thresholds to accurately label the training samples into 'Normal' and 'Distress' classes. Our proposed approach also incorporates data augmentation to address the limited size of the dataset, and the class imbalance. Our model achieved a validation accuracy of 98.3% and a validation loss of 0.2613, reflecting its robustness and capability to generalize effectively. These results highlight the potential that lies in our approach to provide an objective and accurate tool for fetal distress detection.**

I. Introduction

Intrapartum fetal distress refers to the response of a fetus to limitations in oxygen and other factors that contribute to an inadequacy in oxygenation of vital organs, which results in irregularities in fetal heart rate patterns such as tachycardia, bradycardia, accelerations, and decelerations [1]. Similar to the diagnosis of most other physiological conditions or disorders, detecting fetal distress also tends to be a highly subjective task whose accuracy relies on the expertise and experience of healthcare professionals. Fetal distress on its own is a highly complex condition that is determined by the amalgamation of various clinical factors, such as neonatal parameters like ***Apgar* score** [2] [3], which ranges from 0 to 10, calculated as the sum of scores of five neonatal criteria each having a scale of 0 to 2 (as shown in *Table* I), **umbilical arterial *pH*** [4], and maternal risk factors like hypertension, preeclampsia, age, diabetes, etc.

Cardiotocography for fetal health evaluation was first introduced in the mid-20^{th} century [5] by E. Hon [6]. The two main characteristics of the intrapartum CTG are fetal heart rate (FHR) and uterine contractions (UC). The manual interpretation of CTG data involves following the *DR C BRAVADO* [7] approach. This abbreviation summarizes the interpretation; defining the risk (DR), studying the contractions (C), baseline fetal heart rate (BRA) and its variability (V), identifying the presence of accelerations (A), decelerations (D), and lastly the overall assessment of FHR and UC patterns (O). Fetal distress is a condition that is highly complex in nature due to the existence of numerous factors that affect it, and compounding this with the subjective nature of CTG interpretation, there is ample room for inconsistency in the assessment of the same graph among different medical experts. The introduction of CTG for fetal distress assessment is yet to majorly improve prediction rates of abnormal patterns [8]. Hence, it is crucial to work towards developing a highly objective approach to detecting fetal distress, as this could prove to be vital in correctly identifying fetal distress cases, reducing the number of false positives, and ultimately contributing towards reducing infant mortality rates.

II. Literature Review

There are various examples of extensive research and developments that have been made in the field of machine learning-based fetal distress classification. We have examined a range of papers, primarily relating to the identification of the most objective and accurate combination of neonatal outcome measures for data labeling, and also exploring results of past studies to weigh in on the pros and cons of various approaches.

Firstly, we explore a study led by Jun Ogasawara *et al.* [9] where deep learning concepts were applied for fetal distress classification using a dataset of 5406 deliveries from Keio University Hospital. The data was categorized on the basis of *Apgar* score [2] and umbilical arterial *pH* [4], 263 abnormal deliveries were identified. Subsequent screening led to the selection of 162 abnormal and normal records each. The normal cases comprised those with mean umbilical artery

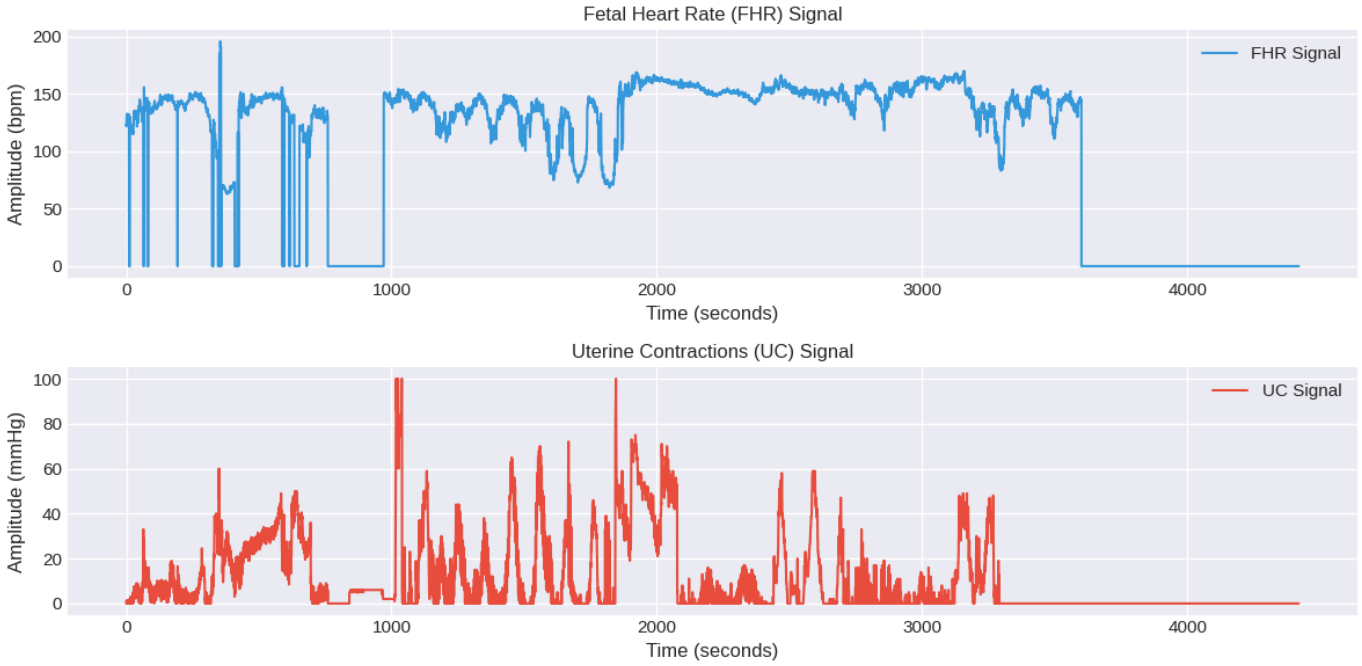

Fig. 1. Fetal Heart Rate and Uterine Contractions of a Distressed Sample *(pH = 7.01, Apgar1 = 2 and Apgar5 = 4)*

pH of 7.31 ± 0.05, while the abnormal cases consisted of samples with a *pH* of 7.25 ± 0.11. *Apgar1* scores in the range 8.31 ± 0.65 were considered normal, whereas scores in the range 4.30 ± 1.81 were considered abnormal cases. The authors implemented CNN-based (CTG-Net) and LSTM-based models, achieving F1 scores of 0.67 ± 0.03 and 0.66 ± 0.04, respectively. These models outperformed other conventional machine learning methodologies such as SVM (with an F1 score of 0.55 ± 0.05) and k-means clustering (with an F1 score of 0.52 ± 0.12).

TABLE I
APGAR SCORE CRITERIA

Sign \ Score	2	1	0
Appearance (skin color)	Pink all over	Pink with Blue extremities	Cyanotic / Pale all over
Pulse (heart rate)	> 100 bpm	< 100 bpm	Absent
Grimace Response (reflex irritability)	Crying, Sneezing, Coughing	Grimacing	No response to stimulation
Activity (muscle tone)	Active	Arms and legs flexed	Absent
Respiration	Good, crying	Slow and irregular, weak, gasping	Absent

In another paper that utilizes a deep learning approach to fetal distress classification by Y. D. Daydulo *et al.* [10], it was highlighted that only samples with a pH of **less than or equal to** 7.15 were labeled as 'Normal' cases. They did not consider *Apgar* score citing that it was a "subjective labeling criteria". Hence, out of the 552 total samples, 439 were labeled 'Normal' and the remaining 113 were 'Distress' samples. This paper delved into the implementation of the classification of distress using Morse wavelet for signal processing coupled with transfer learning with a modified ResNet50 model. They obtained exceptionally high accuracies of 98.7% and 96.1% respectively for two separate stages of labor.

In a study led by H. Liang *et al.* [11], an almost identical labeling approach was used, where samples were labeled 'Distress' only if their pH was only **less than** 7.15. The difference between these two minor variations in pH criteria was that this paper produced 447 'Normal' samples, which is about 8 more 'Normal' samples than the previously reviewed paper [10]. This paper presents the application of 1D CNN coupled with bidirectional GRU to produce another high-performing model that exhibited an accuracy of 96%.

Reviewing another paper by M. O'Sullivan *et al.* [12], we infer that the labeling criteria used was $ph <= 7.0$ and low $(5 <= 6)$ *Apgar5* (5-minute) score as to label the 'Distress' samples. The originally 552-sample CTU-CHB dataset [13] [14] was filtered by removing samples with over 30% traces in the CTG missing. Out of the remaining samples, 310 were labeled as normal samples, 23 were labeled distressed, and 99 samples belonged to neither class. This paper implemented the use of machine learning algorithms such as logistic regression and support vector machines (SVM), with a primary focus on the feature engineering aspect of the CTU-CHB database.

Another study by A. Johny *et al.* [15] highlights the effectiveness of using dynamic learning rates instead of fixed ones. Their experiment involved analysing the use of a variable

learning rate to optimize the accuracy of a CNN model to classify histopathologic (relating to diseased body tissue) image data as 'benign' or 'malignant'. Their results indicated that using a fixed learning rate proves to be counterproductive due to the improvement in accuracy being minimal when the training progresses past 50% of the total epochs. The most optimal cyclic learning rate approach they identified was the use of triangular learning rate, which was first introduced by L.N. Smith [16].

In the paper that originally presents the CTU-CHB database [14] by V. Chudáček *et al.* [13], *Table 2* showcases over 21 papers relating to the various distress classification criteria that were employed on numerous databases. Among these, 13 papers used pH as a criterion for the classification of the normalcy of a sample, and 5 papers used the next most frequently occurring criterion, which was *Apgar* score. Taking into account the results of past research and the promising results obtained in the papers that have been cited above, we have also decided to implement the application of deep learning and focus on using a similar set of criterion to develop an optimal model and ultimately improve the detection of fetal distress based on cardiotocography data.

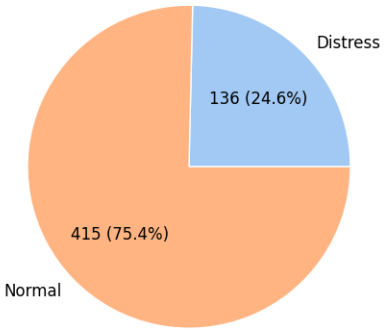

Fig. 2. Distribution of Classes after Labeling

III. Dataset Description

The CTU-CHB Intrapartum Cardiotocography Database v1.0.0 [13] includes a set of 552 carefully selected samples of CTG recordings from a total of 9164 recordings. Each CTG has been recorded over a 90-minute duration, starting at most 90 minutes prior to the delivery of a child. Most of the recordings deliberately comprise of vaginal deliveries with only 46 sections included from cesarean deliveries. Each sample comprises of two files, a *.dat* and a *.hea* file. The *.dat* file contains the CTG recording's signal data, as visualized in *Fig.* 1. On the other hand, the *.hea* is a header file that contains all metadata of the signals recording specifications of various physiological parameters such as fetal outcome measures (pH, base excess, base deficit, pCO_2, *Apgar1* and *Apgar5* scores), maternal risk factors (age, presence of comorbidities such as gestational diabetes, hypertension, and preeclampsia, presence of meconium-stained fluid and maternal fever, etc.), fetus descriptors (such as the gestational period in weeks, weight and sex), delivery descriptors and also signal information such as sample rate, number of recording channels, etc. The labeling of the samples is also done by comparing the parameter values (pH and *Apgar*) extracted from these header files with the thresholds we have defined.

IV. Methodology

Deep learning models generally tend to outperform conventional machine learning models in most scenarios, especially when dealing with time series data such as FHR and UC signals. Since our proposed approach uses convolutional neural networks, we had to first augment the dataset and scale it up to a sufficiently large number of samples. This is considering the fact that CNNs require a substantial amount data to learn from, to generalize effectively well and produce a reliable classification model that is also not overfitting.

In our proposed methodology, we are required to first label the samples in the dataset, extract and preprocess the signals stored in the *.dat* files, perform data augmentation, and finally train the CNN model with various suitable hyperparameters and other optimization techniques to yield the best performing model possible. Let us delve into the specifics of every stage in the proposed implementation.

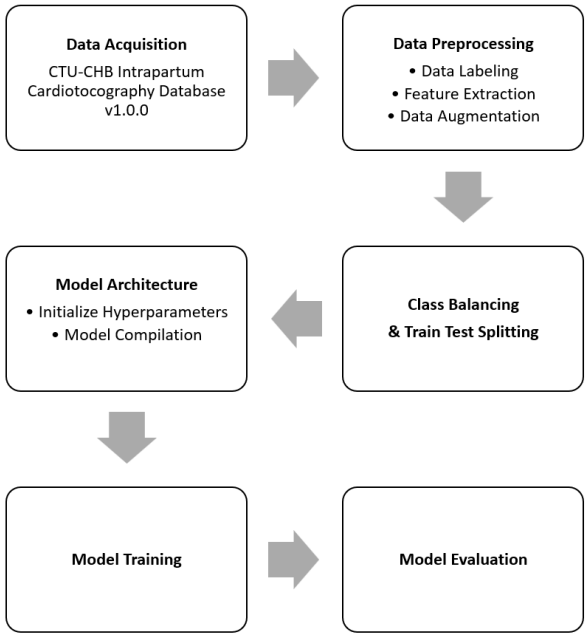

Fig. 3. Architecture of Proposed Classification Methodology

A. Data Labeling

Since the CTU-CHB database is not already labeled, the first step was the process of assigning labels to each of the 552 samples. We identified and decided to use the most significant parameters that accurately indicate fetal distress for the categorization of the unlabeled dataset. Hence, we first read the contents of the *.hea* file and compared each sample's umbilical arterial pH level and *Apgar* scores at 1 and 5 minutes after birth with our predefined thresholds. The

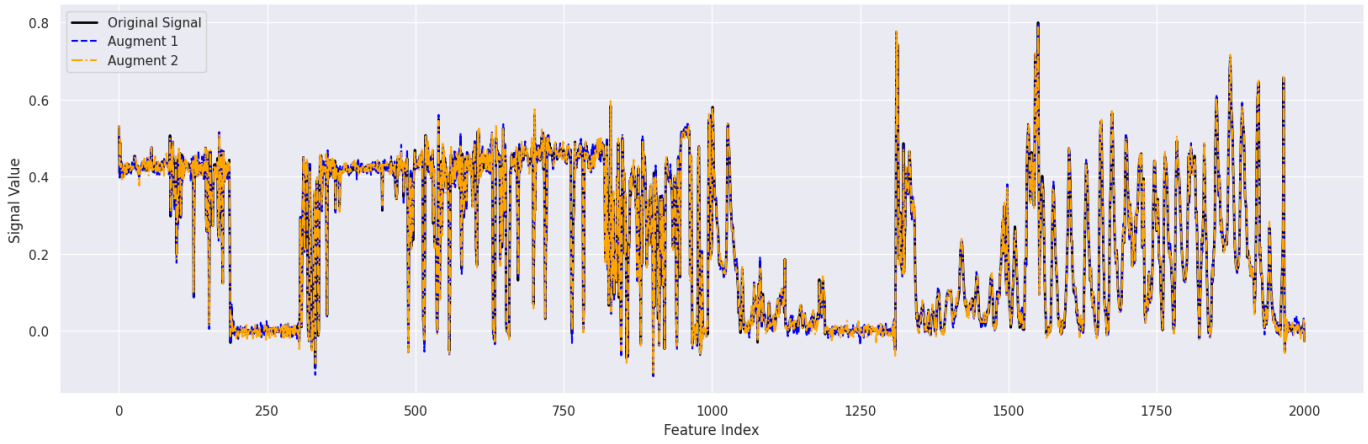

Fig. 4. Comparison of preprocessed FHR and UC features of a sample with its two most distinctly augmented samples.

distress thresholds that were defined are as follows: If the *pH* level falls below 7.15 $\mathbf{(pH < 7.15)}$ or the Apgar scores at 1 or 5 minutes are less than 7 $\mathbf{(Apgar1/Apgar5 < 7)}$, the corresponding CTG recording is labeled as 'Distress'. Otherwise, the CTG recording is labeled as 'Normal'. We had previously established the justification of this labeling strategy, *i.e.* capturing instances where the **most critical** physiological markers are indicating fetal distress. This labeling criteria resulted in the following class distribution of data, with 415 samples labeled as 'Normal' and the remaining 136 samples labeled as 'Distress' as shown in *Fig.* 2

B. Data Preprocessing

In the preprocessing phase, the Fetal Heart Rate (FHR) and Uterine Contraction (UC) signals are extracted from the CTU-CHB database using the **wfdb** Python library. The raw signals of each sample are extracted, and the FHR and UC components are isolated based on their corresponding signal names. Since we are dealing with time series data, we resampled the extracted FHR and UC signals down to 1000 features each to ensure consistency in the length of all signals.

FHR and UC are differently measured quantities having different units. FHR is measured in beats per minute whereas UC is essentially uterine pressure, which is measured in *mmHg*. Hence, we had to normalize these two signals to avoid any feature dominance and ensure homogeneity of data being used to train the model. For this, we used Min-Max normalization to the resampled Fetal Heart Rate and Uterine Contraction signals. This normalization involves dividing each value in the FHR signal by the maximum FHR value ($\max_{\text{FHR}}$), which was 293.0, and each value in the UC signal by the maximum UC value ($\max_{\text{UC}}$), which was 127.5. Hence, all values were scaled to a uniform range of $[0, 1]$.

This process is repeated for every sample, and after all samples' FHR and UC features are normalized and concatenated into 2000-feature one-dimensional arrays, they were all appended to a Python list that stores the preprocessed signals of all 552 samples cumulatively. This results in a list with dimensions (552x2000), ensuring that the dataset is suitable for further analysis and model training.

C. Data Augmentation

As previously discussed, a major concern with the CTU-CHB database [13] [14] is the relatively small size of the dataset. With only 552 samples, training deep neural network models like CNNs is difficult. Hence, we decided to implement data augmentation by artificially increasing the number of samples using the existing samples. We utilized the **tsaug** Python library for the augmentation of the resampled signals, and specifically applied the **AddNoise** transformation with a scaling factor of 0.01. This transformation introduces jittering to the original signals, which essentially adds subtle variations to the signal, as shown in *Fig.* 4. This augmentation transformation is applied with a 90% probability on each sample. Since the goal was to build a large enough dataset, we ensured that for each sample, 20 augmented samples were generated.

Lastly, we further reshaped the augmented dataset by vertically stacking the original and augmented data, and shuffled the augmented dataset to increase variability and sparsity in the dataset before training. This prevents potential biases before the model is trained. After augmentation, we were able to

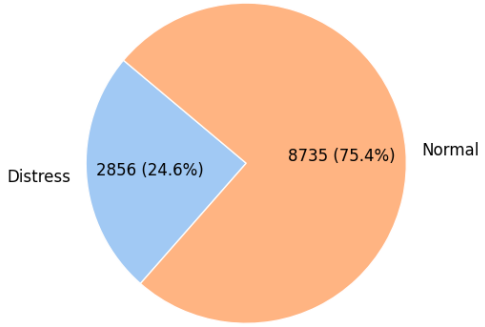

Fig. 5. Distribution of Classes (after Augmentation)

increase the total number of samples from 552 to 11592, which included the original 552 samples. The class distribution after augmentation is shown in *Fig.* 5

D. Data Balancing, Splitting & Further Preprocessing

While we had successfully augmented and expanded the size of the dataset, the severe imbalance in class distribution remained. As shown in *Fig.* 5, the number of 'Distress' samples is only 2856, which is significantly less than the 8735 'Normal' classes. To balance the dataset before splitting it into training and testing sets, we decided to select 2856 random 'Normal' samples and ensure there was a 50:50 class distribution, similar to the approach taken by H. Liang *et al.* [11] This would ensure that the model training would be effective and yield better validation results too. Once we had a balanced dataset of 5712 samples, we split it into training and testing sets in a 70:30 ratio. Hence, we finally had 3998 training and 1714 validation samples to train and evaluate the CNN model, respectively.

We encoded the categorical 'Normal' and 'Distress' labels into numerical values to provide and ensure the data is suitable as input to the CNN model. We also reshaped the train and test sets by adding a third dimension to convert the previously two-dimensional matrix to a three-dimensional tensor input.

TABLE II
MODEL ARCHITECTURE

Layer	Component	Output Shape
0	Input (Conv1D + ReLU)	(None, 1998, 32)
1	MaxPooling1D	(None, 999, 32)
2	BatchNormalization	(None, 999, 32)
3	Dropout	(None, 999, 32)
4	Conv1D + ReLU	(None, 997, 64)
5	MaxPooling1D	(None, 498, 64)
6	BatchNormalization	(None, 498, 64)
7	Dropout	(None, 498, 64)
8	Flatten	(None, 31872)
9	Dense + ReLU	(None, 64)
10	Dropout	(None, 64)
11	Dense + Sigmoid	(None, 1)
Total params		2,046,657
Trainable params		2,046,465
Non-trainable params		192

E. Model Architecture and Hyperparameters

The architecture consists of a set of 32 and 64-filter one-dimensional convolutional layers of size 3, coupled with rectified linear unit (ReLU) activation function, which is effective in learning intricate features. We also added one-dimensional max pooling layers of window size 2 to downsample the spatial dimensions. Batch normalization enhances training stability and accelerates convergence coupled with dropout layers to prevent overfitting by randomly deactivating neurons during training. The dropout was set to 0.5 after each batch normalization, which means 50% of neurons are dropped. After flattening the output into a one-dimensional vector, **L2 regularization** was also performed with a penalty of 0.01 to the kernel weights, to further prevent overfitting. Meanwhile, a third dropout of 0.7 (70%) was applied after the first fully connected layer. The model was compiled with Adam optimizer with a custom dynamic learning rate configuration and **binary cross-entropy** loss function. *Table* II summarizes this model architecture and the output shapes after each layer. The model was trained for 100 epochs, with a **batch size** of 64 and an **initial learning rate** of $1.11\text{x}10^{-4}$, using a **dynamic** learning rate by implementing TensorFlow's callback function. It was configured to monitor the **validation loss** and adjust the learning rate by a factor of 0.5 if no improvement is observed for two consecutive epochs. This reinforces the adaptive optimization of model performance.

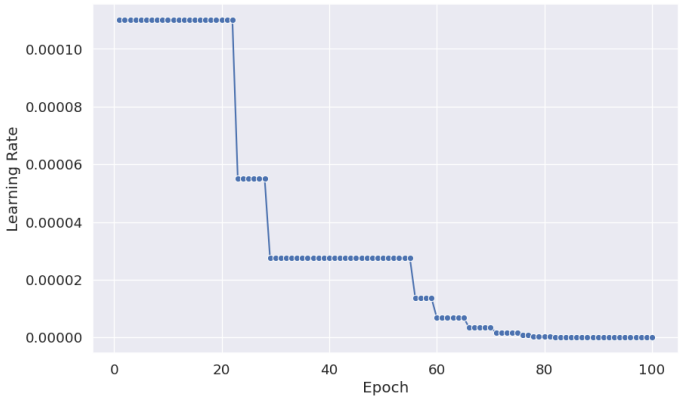

Fig. 6. Learning Rate over Epochs

V. RESULTS

The proposed CNN model exhibited promising results in our fetal distress classification task. A major reason for concatenating the resampled FHR and UC signal features was to qualitatively study and understand if an inherent correlation exists between them, and in turn identifying if the model can capture it to produce an accurate classification of a sample. To justify this, we obtained an exceptionally high validation accuracy of 98.3%, and a smoothly converging model with a final validation loss of 0.2613 (*Fig.* 7, 8). The training accuracy

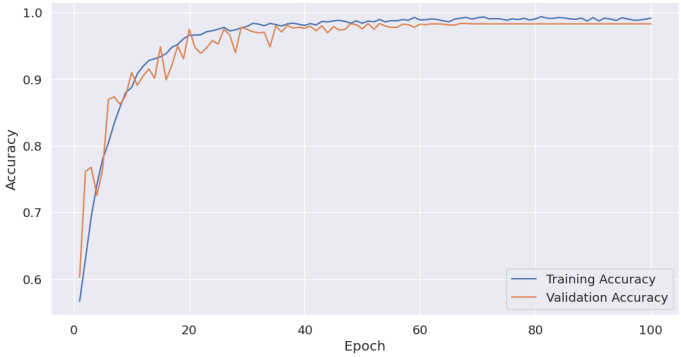

Fig. 7. Training and Validation Accuracy

steadily increased throughout the epochs, and plateaued at around 99%, while never reaching 100%, hence, showing no signs of overfitting. The training and validation losses followed a similar trend, and the consistency between the respective losses also suggests that the model is not overfitting. The sensitivity of the 'Distress' class obtained was 97%, while it was 100% for the 'Normal' class.

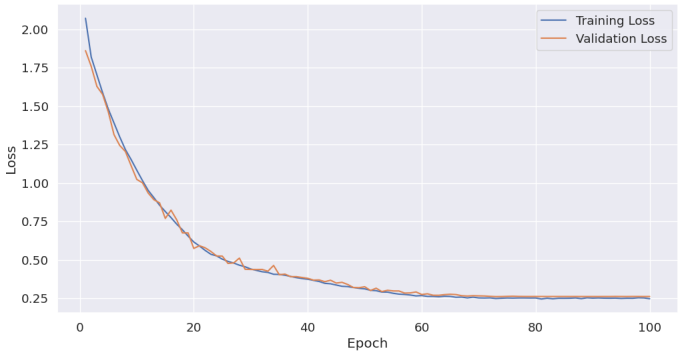

Fig. 8. Training and Validation Loss

On analyzing the confusion matrix (*Fig* 9) of the model on the validation data (1714 samples), we can also conclude that the model generalized well to unseen validation data. These results reflect the robustness and consistency of the model that we have developed.

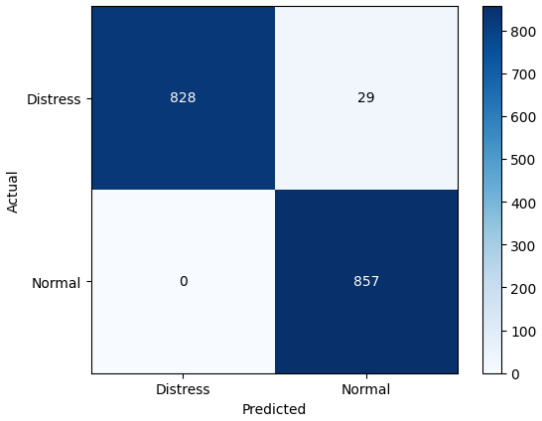

Fig. 9. Confusion Matrix

VI. Discussion and Conclusion

In conclusion, the developed one-dimensional CNN model presents a promising approach to intrapartum fetal distress classification using FHR and UC signals. The high validation accuracy and minimal loss indicates that the model could be a valuable tool for healthcare professionals to accurately identify cases of fetal distress. We believe that this research strongly contributes to the ongoing efforts in utilizing machine learning approaches, specifically deep learning, for improving fetal distress detection. The objectivity introduced by the model reduces the subjectivity associated with the manual interpretation of CTG data.

While the results are promising, there is a lot of scope for development from an applicative standpoint, and we hope to explore various new aspects in future work, including clinical validation of the results of our model to eventually implement it as a tool for real-time use. Our future aspirations include building software that uses our CNN model to assist medical professionals in objectively validating their CTG interpretations, and exploring improved methods of preprocessing the CTG signal data and advanced methodologies of data augmentation. In summary, we believe that this work has the potential to impact clinical practices and contribute to significant advancements in perinatal care.

References

[1] J. Parer and E. Livingston, "What is fetal distress?" *American journal of obstetrics and gynecology*, vol. 162, no. 6, pp. 1421–1427, 1990.

[2] V. Apgar, "A proposal for a new method of evaluation of the newborn infant." *Anesthesia & Analgesia*, vol. 32, no. 4, pp. 260–267, 1953.

[3] L. V. Simon, M. F. Hashmi, and B. N. Bragg, "Apgar score," 2017.

[4] L. S. James, I. Weisbrot, C. Prince, D. Holaday, and V. Apgar, "The acid-base status of human infants in relation to birth asphyxia and the onset of respiration," *The Journal of pediatrics*, vol. 52, no. 4, pp. 379–394, 1958.

[5] D. Ayres-de Campos, "Electronic fetal monitoring or cardiotocography, 50 years later: what's in a name?" *American Journal of Obstetrics & Gynecology*, vol. 218, no. 6, pp. 545–546, 2018.

[6] E. H. Hon, "The electronic evaluation of the fetal heart rate: Preliminary report," *American Journal of Obstetrics & Gynecology*, vol. 75, no. 6, pp. 1215–1230, 1958.

[7] R. E. Bailey, "Intrapartum fetal monitoring," *American family physician*, vol. 80, no. 12, pp. 1388–1396, 2009.

[8] J. A. Spencer, "Clinical overview of cardiotocography," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 100, pp. 4–7, 1993.

[9] J. Ogasawara, S. Ikenoue, H. Yamamoto, M. Sato, Y. Kasuga, Y. Mitsukura, Y. Ikegaya, M. Yasui, M. Tanaka, and D. Ochiai, "Deep neural network-based classification of cardiotocograms outperformed conventional algorithms," *Scientific reports*, vol. 11, no. 1, p. 13367, 2021.

[10] Y. D. Daydulo, B. L. Thamineni, H. K. Dasari, and G. T. Aboye, "Deep learning based fetal distress detection from time frequency representation of cardiotocogram signal using morse wavelet: research study," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 329, 2022.

[11] H. Liang, Y. Lu, Q. Liu, and X. Fu, "Fully automatic classification of cardiotocographic signals with 1d-cnn and bi-directional gru," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 4590–4594.

[12] M. O'Sullivan, T. Gabruseva, G. B. Boylan, M. O'Riordan, G. Lightbody, and W. Marnane, "Classification of fetal compromise during labour: signal processing and feature engineering of the cardiotocograph," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1331–1335.

[13] V. Chudáček, J. Spilka, M. Burša, P. Jank, L. Hruban, M. Huptych, and L. Lhotská, "Open access intrapartum ctg database," *BMC pregnancy and childbirth*, vol. 14, pp. 1–12, 2014.

[14] A. L. Goldberger *et al.*, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation [Online]*, vol. 101, no. 23, pp. e215–e220, 2000.

[15] A. Johny and K. Madhusoodanan, "Dynamic learning rate in deep cnn model for metastasis detection and classification of histopathology images," *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–13, 2021.

[16] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.