Advancing Image Classification with Explainable AI: From Fetal Health to Fashion Using CNN and Grad-CAM

A Course Application Report

submitted as part of the course

Explainable Artificial Intelligence

BCSE418L

School Of Computer Science and Engineering

VIT Chennai

Winter 2023-2024

Course Faculty : Dr. B Radhika Selvamani @miscVIT Chennai - Learning Management Systems, url=https://lms.vit.ac.in/, journal=VIT Chennai - Learning Management Systems []

Submitted By

Sushmetha S R	21BAI1162
Pritika Kannapiran	21BAI1172
Dhanush M	21BAI1744

ABSTRACT

This study explores the interpretability of machine learning models using GradCam and SHAP techniques applied to two distinct datasets. First, the Fashion MNIST dataset is used to explain a basic convolutional neural network (CNN) model. GradCam, a gradient-based visualization method, is employed to highlight the regions of interest within the images that contribute most significantly to the model's predictions. Additionally, SHAP (SHapley Additive exPlanations) is utilized to provide insights into feature importance and decision-making processes of the CNN.

Furthermore, the study extends its analysis to a fetal health classification dataset, employing a random forest model. Here, SHAP is applied to elucidate the random forest's decision mechanism and identify the key features influencing the prediction outcomes related to fetal health. By leveraging these interpretability techniques, the study aims to enhance the transparency and trustworthiness of machine learning models, especially in critical domains like healthcare. The findings underscore the utility of GradCam and SHAP in uncovering model behaviors and improving the understanding of complex datasets and model predictions.

Contents

INTRODUCTION

In the realm of machine learning, the complexity and opacity of models often pose significant challenges to their adoption and trustworthiness, particularly in critical applications like healthcare. This study addresses the fundamental issue of model interpretability by leveraging advanced techniques such as GradCam and SHAP (SHapley Additive exPlanations) on two distinct datasets: the Fashion MNIST dataset and a fetal health classification dataset.

The overarching goal of this research is to demystify the decision-making processes of machine learning models. Specifically, we aim to shed light on the inner workings of a convolutional neural network (CNN) applied to the Fashion MNIST dataset and a random forest model applied to fetal health classification. By utilizing GradCam, which highlights significant image regions for CNN-based predictions, and SHAP, which quantifies feature importance in predictive models, this study seeks to elucidate how these models arrive at their conclusions.

Real-world users, including healthcare practitioners, researchers, and policymakers, stand to benefit greatly from transparent machine learning models. These users rely on accurate and interpretable predictions for informed decision-making in areas like medical diagnosis and prognosis. In healthcare, understanding the rationale behind a model's predictions is crucial for building trust and facilitating the adoption of machine learning technologies.

By enhancing model interpretability through methodologies like GradCam and SHAP, this research contributes to fostering trust in machine learning models, ultimately paving the way for their responsible and effective deployment in critical applications. The insights gained from this study can inform the development of more interpretable and reliable machine learning systems, benefiting both practitioners and end-users in various domains, including healthcare and beyond.

RELATED WORKS

[1] Bartler, A., Hinderer, D., & Yang, B. (2021). Grad-LAM: Visualization of Deep Neural Networks for Unsupervised Learning. In *2020 28th European Signal Processing Conference (EUSIPCO)* (pp. 1407-1411). Amsterdam, Netherlands. doi: 10.23919/Eusipco47968.2020.9287730.

[2] Kim, Y., & Kim, Y. (2022). Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models. *Sustainable Cities and Society*, 79, 103677. https://doi.org/10.1016/j.scs.2022.103677.

[3]A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, Art. no. 6, May 2017, doi: 10.1145/3065386.

DESIGN

1. Fetal Health Classification:

- **Exploring XAI Methods in Fetal Health Classification**

  Our project adopts a systematic approach to enhance fetal health assessment using Explainable Artificial Intelligence (XAI) techniques. Central to our methodology is the utilization of the Random Forest classifier, a powerful ensemble learning algorithm known for its robustness and interpretability. We integrate state-of-the-art XAI methods to gain insights into the decision-making process of our model, thereby augmenting the interpretability and transparency of fetal health predictions.

- **Leveraging the Power of Random Forest Classifier**

  The Random Forest classifier serves as the cornerstone of our predictive model. Its ability to handle complex datasets, capture non-linear relationships, and mitigate overfitting aligns perfectly with the intricacies of fetal health data. By leveraging an ensemble of decision trees, the Random Forest model achieves high accuracy while maintaining interpretability, making it an ideal choice for our XAI-driven approach.

- **Unveiling Feature Importance with SHAP Values**

  We employ the Shapley Additive Explanations (SHAP) method to unravel the importance of features contributing to our model's predictions. SHAP values provide a comprehensive understanding of each feature's impact, enabling healthcare professionals to prioritize critical factors in fetal health assessment. Through visualizations such as beeswarm plots and summary plots, we elucidate the significance of features like fetal health, uterine contractions, and accelerations, empowering users to make informed decisions based on actionable insights.

- **Visualizing Decision Trees for Enhanced Interpretability**

  In addition to SHAP values, we harness the interpretability of decision trees within the Random Forest ensemble. Visual representations of decision trees elucidate the hierarchical decision-making process, offering clinicians and researchers a transparent view of how the model classifies fetal health states. This transparency not only enhances trust in the model but also facilitates collaboration between AI systems and domain experts in obstetrics.

- **Advancing XAI for Improved Healthcare Outcomes**

  By combining advanced XAI methodologies with the Random Forest model, our project propels the field of fetal health assessment towards greater transparency, interpretability, and actionable insights. Our design emphasizes the synergy between machine learning algorithms and XAI techniques, paving the way for enhanced decision-making in healthcare domains critical for maternal and infant well-being.

2. Fashion MNIST Classification:

Our project aims to develop an Explainable AI (XAI) framework for enhancing the interpretability and transparency of deep learning models used in fashion image classification, specifically focusing on the Fashion-MNIST dataset. The design of our XAI framework encompasses various methodologies and techniques to provide insightful explanations for model predictions, aiding in understanding and improving the model's performance.

- **Deep Learning Model: Convolutional Neural Network (CNN)**
  - We leverage a CNN architecture optimized for image classification tasks, utilizing TensorFlow and Keras libraries.
  - The CNN model is trained on the Fashion-MNIST dataset, comprising grayscale images of fashion items categorized into ten classes.
  - Model training includes data preprocessing, normalization, and validation techniques to ensure robustness and generalization.

- **XAI Methods and Techniques:**

  a. SHAP (SHapley Additive exPlanations):

  - We integrate SHAP values to explain the contributions of individual pixels/features to the model's predictions.
  - SHAP values provide insights into feature importance and help identify critical regions in input images that influence classification decisions.

b. GradCAM (Gradient-weighted Class Activation Mapping):

- GradCAM generates class activation maps highlighting image regions crucial for predicting specific classes.
- By visualizing GradCAM outputs, we gain interpretability into the model's attention mechanism, understanding which parts of the image drive its predictions.

- **Data Analysis and Preprocessing:**
  - We conduct exploratory data analysis (EDA) to understand class distributions, pixel ratios, and image characteristics in the Fashion-MNIST dataset.

  - Preprocessing steps include data normalization, reshaping, and partitioning into training and testing sets for model evaluation.

- **Model Training and Evaluation:**
  - The CNN model is trained using an Adam optimizer, sparse categorical cross-entropy loss, and accuracy metrics.

  - We evaluate model performance through metrics such as accuracy, confusion matrix analysis, and classification reports, ensuring a comprehensive assessment of classification capabilities.

- **Explanation and Visualization:**
  - Using SHAP and GradCAM, we generate visual explanations for model predictions, showcasing influential features and image regions for each class.

  - Visualization techniques include image plots, heatmap overlays, and comparative analyses between correct and incorrect predictions, enhancing interpretability and trust in model decisions.

EXPLORATORY DATA ANALYSIS

Figure 1 illustrates a dataset comprising 10 distinct classes, with each class containing 6000 samples for the training set and 1000 samples for the testing set. This indicates a balanced distribution of image samples across all classes. Prior to conducting in-depth Exploratory Data Analysis (EDA), images were normalized by scaling their values to a range of 0 to 1 in both datasets. Our comprehensive EDA consists of two main components: a statistical analysis of instance sizes across classes and a visualization technique utilizing dimensionality reduction to highlight intra-class feature similarities.

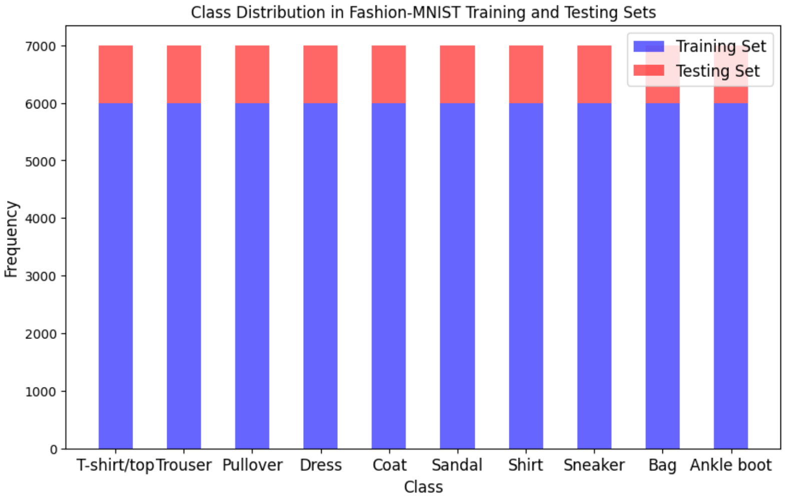


Fig. 1. Class frequency in Fashion-MNIST training and testing sets

In our analysis, we investigate the distribution of pixel ratios for each image per class in both the training and testing sets, recognizing that the variance and size of instance pixels are crucial indicators of dataset complexity and challenges. The ratio of non-background pixels (instance pixels) is calculated for each image across both datasets. To identify statistical differences per class between the datasets, a one-way ANOVA test is conducted. The results of these tests are visually depicted in Figure 2, presenting the ratio distributions for each dataset. Notably, classes such as 'sneaker' exhibit smaller mean and variance in pixel ratios, while 'sandal' and 'bag' display larger variances. A significant observation, as outlined in Figure 3, is the statistical disparity in the instance size distribution of the 'coat' class between the two datasets (P-value=0.006). This finding suggests potential imbalances in instance size distributions, indicating an area that may warrant further investigation to understand dataset representativeness and model training effectiveness.

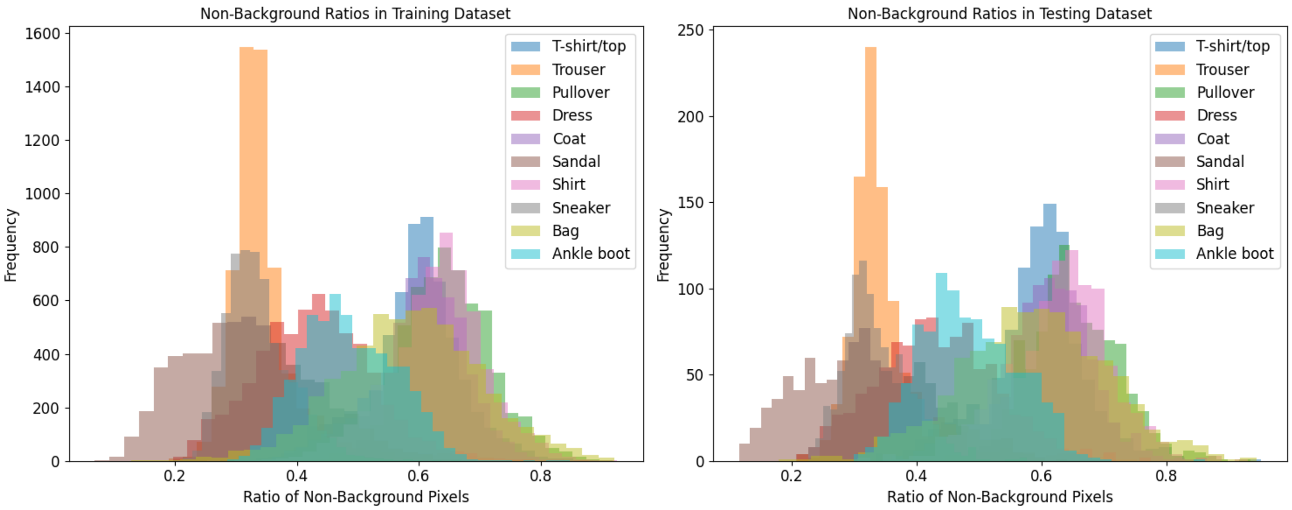
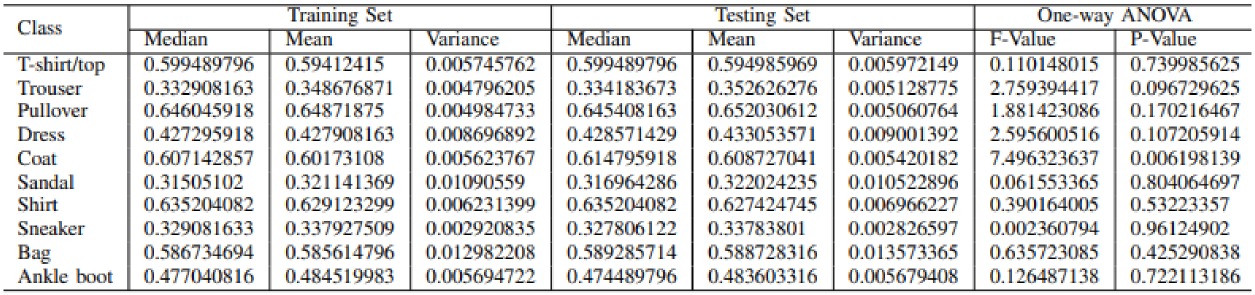


Fig. 2. Non-background ratios in Fashion-MNIST training and testing sets



Class	Training Set			Testing Set			One-way ANOVA	
	Median	Mean	Variance	Median	Mean	Variance	F-Value	P-Value
T-shirt/top	0.599489796	0.59412415	0.005745762	0.599489796	0.594985969	0.005972149	0.110148015	0.739985625
Trouser	0.332908163	0.348676871	0.004796205	0.334183673	0.352626276	0.005128775	2.759394417	0.096729625
Pullover	0.646045918	0.64871875	0.004984733	0.645408163	0.652030612	0.005060764	1.881423086	0.170216467
Dress	0.427295918	0.427908163	0.008696892	0.428571429	0.433053571	0.009001392	2.595600516	0.107205914
Coat	0.607142857	0.60173108	0.005623767	0.614795918	0.608727041	0.005420182	7.496323637	0.006198139
Sandal	0.31505102	0.321141369	0.01090559	0.316964286	0.322024235	0.010522896	0.061553365	0.804064697
Shirt	0.635204082	0.629123299	0.006231399	0.635204082	0.627424745	0.006966227	0.390164005	0.53223357
Sneaker	0.329081633	0.337927509	0.002920835	0.327806122	0.33783801	0.002826597	0.002360794	0.96124902
Bag	0.586734694	0.585614796	0.012982208	0.589285714	0.588728316	0.013573365	0.635723085	0.425290838
Ankle boot	0.477040816	0.484519983	0.005694722	0.474489796	0.483603316	0.005679408	0.126487138	0.722113186

Fig 3 Descriptive analysis and one-way ANOVA comparing instance pixel ratios between Fashion-MNIST training and testing sets.

Subsequently, we utilize the Uniform Manifold Approximation and Projection (UMAP) technique to visually represent intra-class feature similarities within the training set. As shown in Figure 4, this method maps dataset features onto a reduced dimensional space, where the proximity between points reflects the degree of similarity between different classes. Noteworthy similarities are observed between classes such as 'sneaker' and 'sandal', as well as between 'dress' and 't-shirt/top', and 'coat' and 'pullover', revealing consistent patterns and relationships within the data.

The heatmap shows the correlation between the different features in the dataset. The values range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. The heatmap shows that there are several features that are highly correlated with each other. For example, the features "baseline value" and "accelerations" are highly positively correlated, which means that as the value of one feature increases, the value of the other feature also tends to increase. Similarly, the features "uterine contractions" and "fetal_movement" are highly negatively correlated, which means that as the value of one feature increases, the value of the other feature tends to decrease. These correlations can be useful for feature selection and model building.

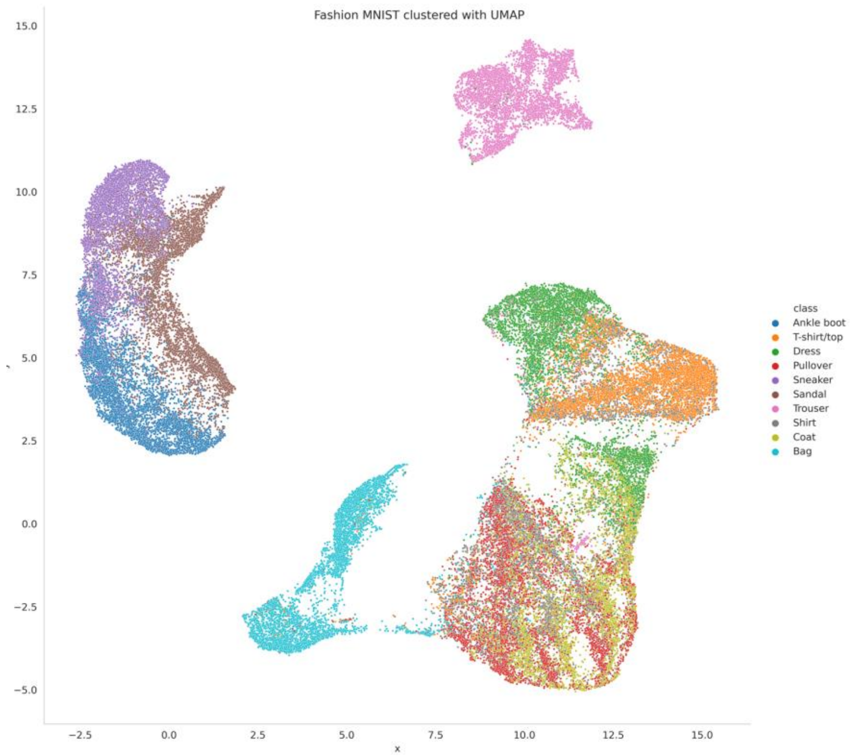


Fig. 4. Dimension reduction using UMAP on Fashion MNIST

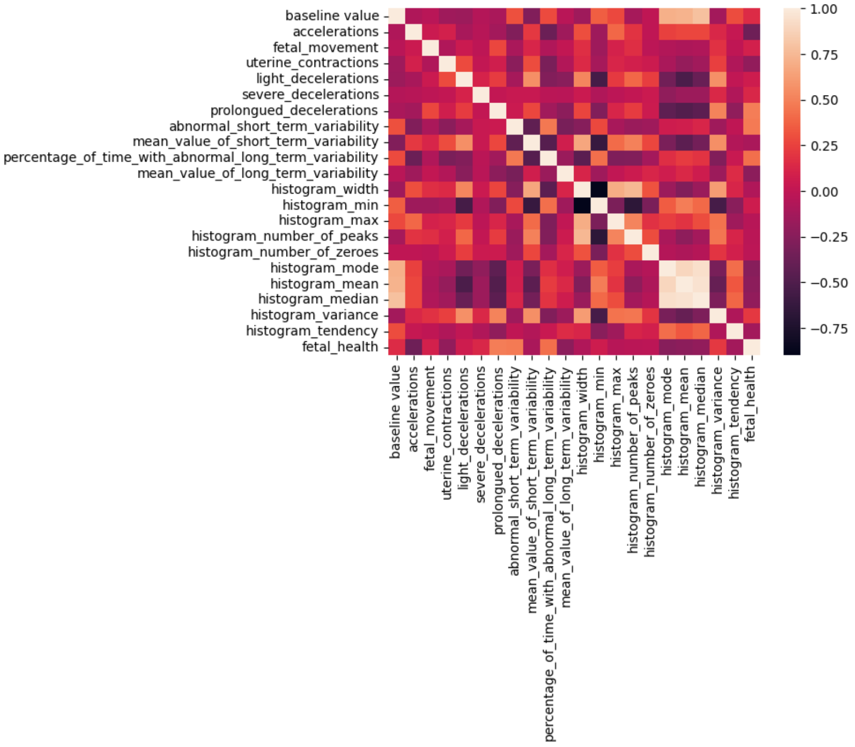


Fig 5 Heatmap depicting the correlations between the independent variables contributing to fetal health

RESULTS AND DISCUSSION

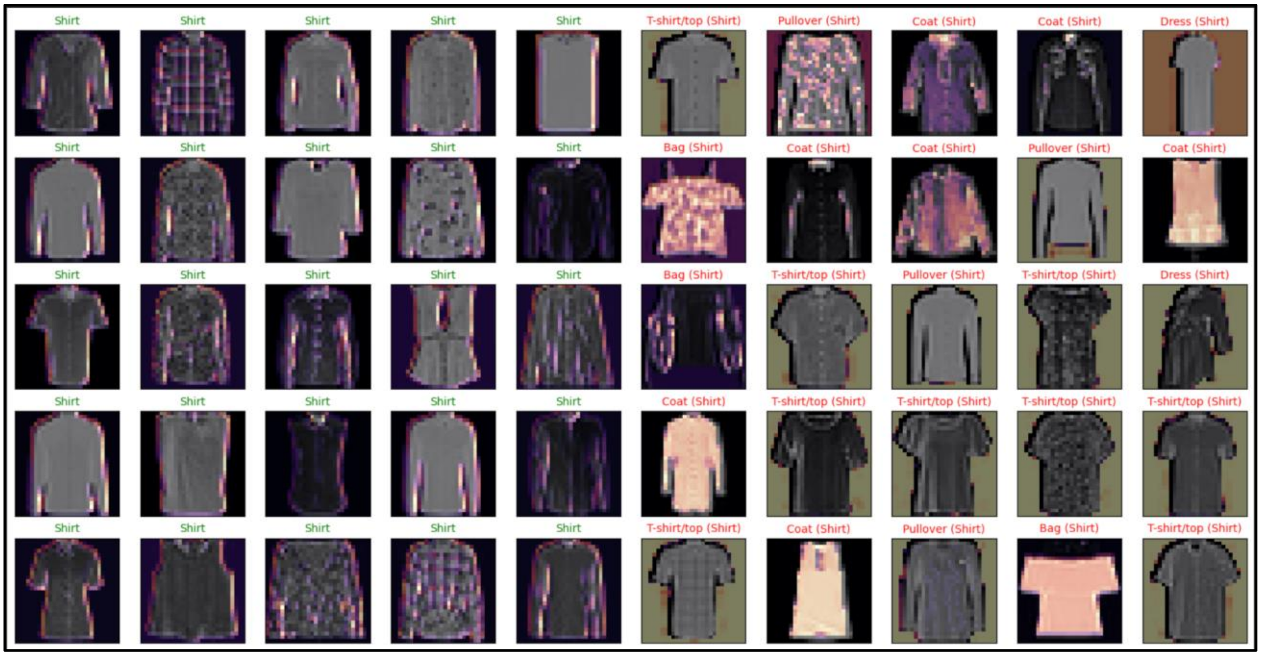
This research delves into enhancing the interpretability of machine learning models through the application of GradCam and SHAP techniques on two distinct datasets. Initially, we examine the Fashion MNIST dataset using a straightforward convolutional model. GradCam is utilized to visually identify and emphasize the specific regions within fashion images that have the most significant impact on the model's predictions. Additionally, SHAP is employed to clarify the significance of various features in the classification process, providing valuable insights into the decision-making process of the model.

Furthermore, SHAP analysis is extended to a fetal health classification dataset, employing a random forest model. In this context, SHAP aids in identifying the critical features that influence the model's predictions regarding fetal health status. Through the application of these interpretability techniques, this study enhances transparency and comprehension of model behaviors across various datasets and architectures.

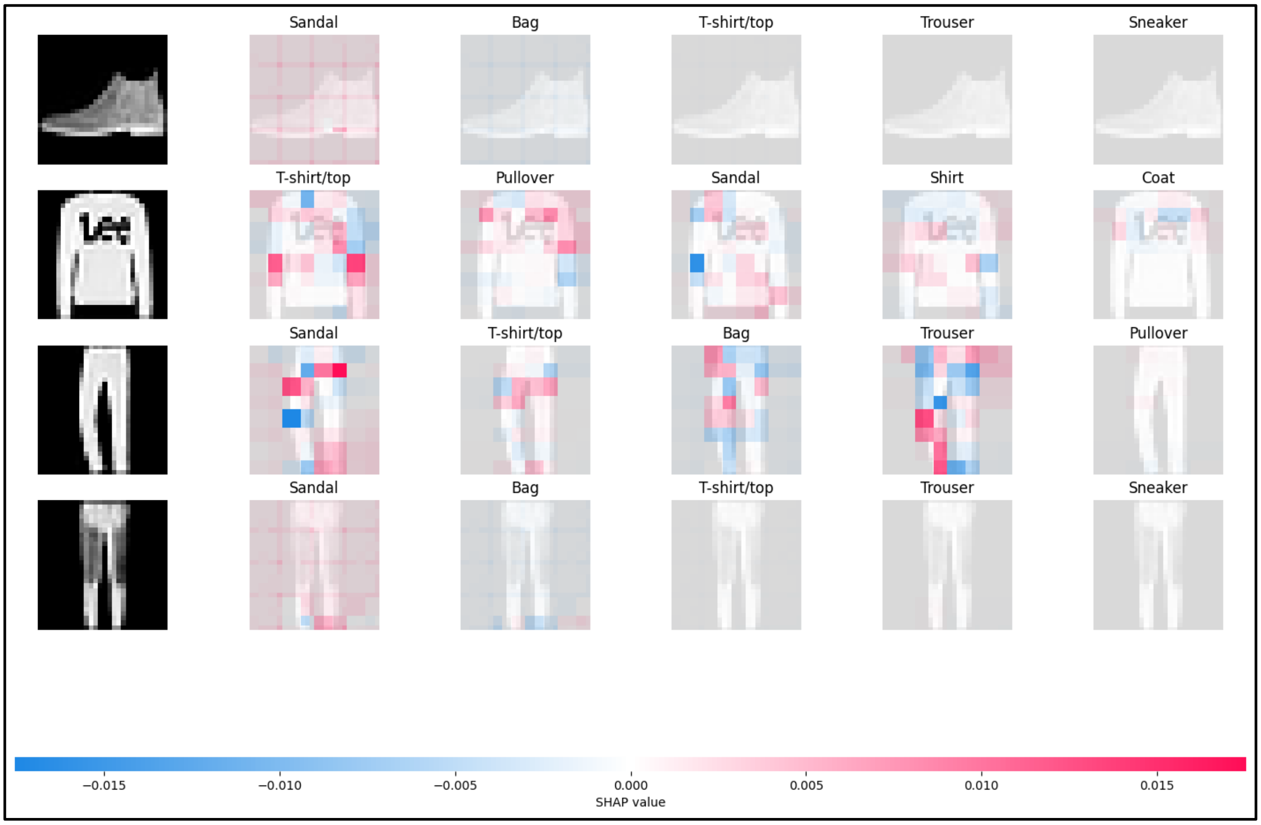
The discoveries from this research contribute to the broader objective of improving machine learning model interpretability, particularly in vital domains like healthcare. By enabling practitioners and researchers to gain deeper insights into model predictions, these techniques facilitate more informed decision-making and cultivate trust in the reliability and practicality of machine learning algorithms.

Fashion MNIST:

GradCAM visualization

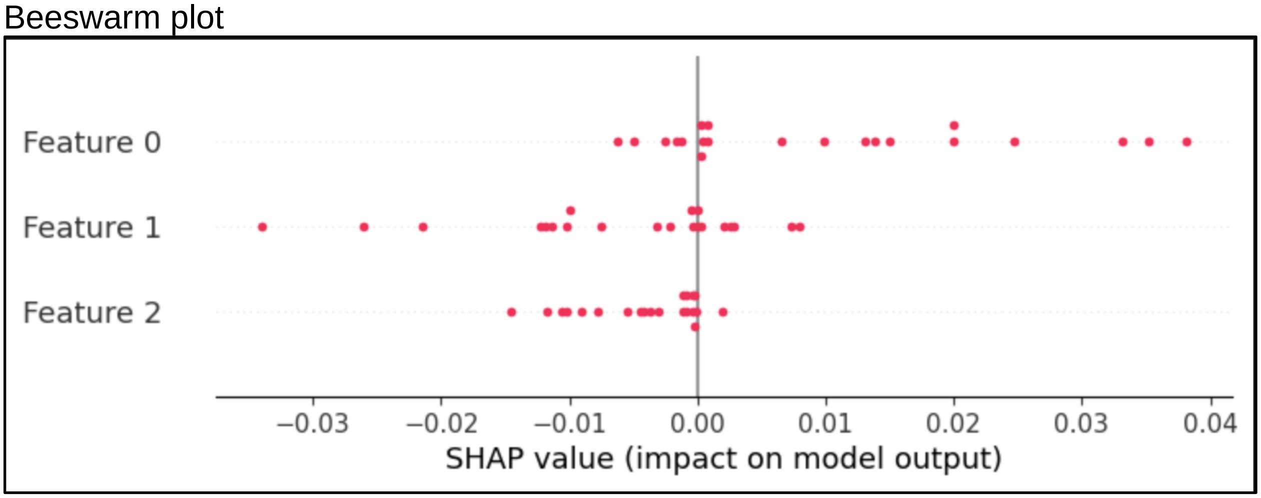


Partition Explainer (SHAP) visualization

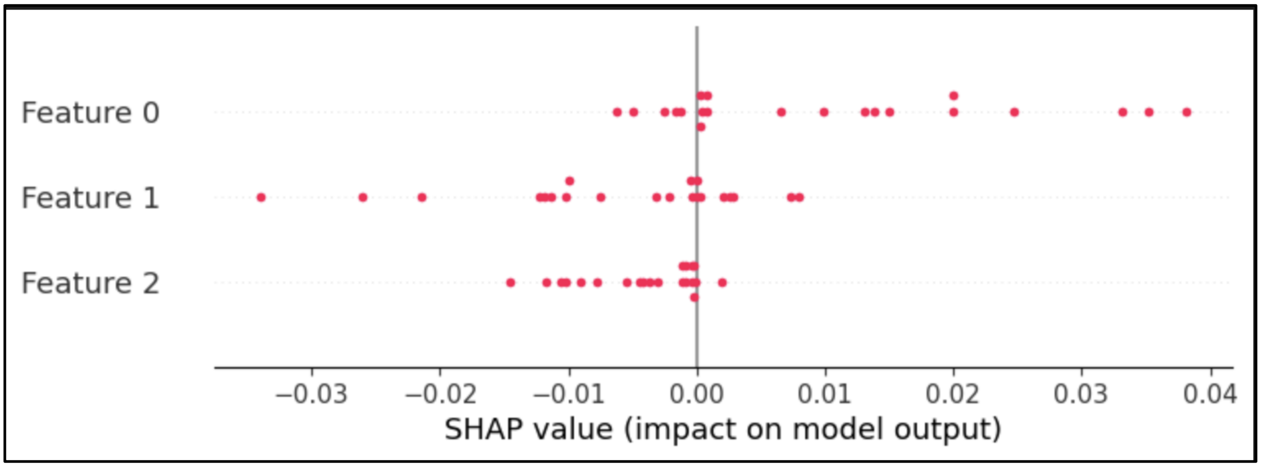


Fetal Health

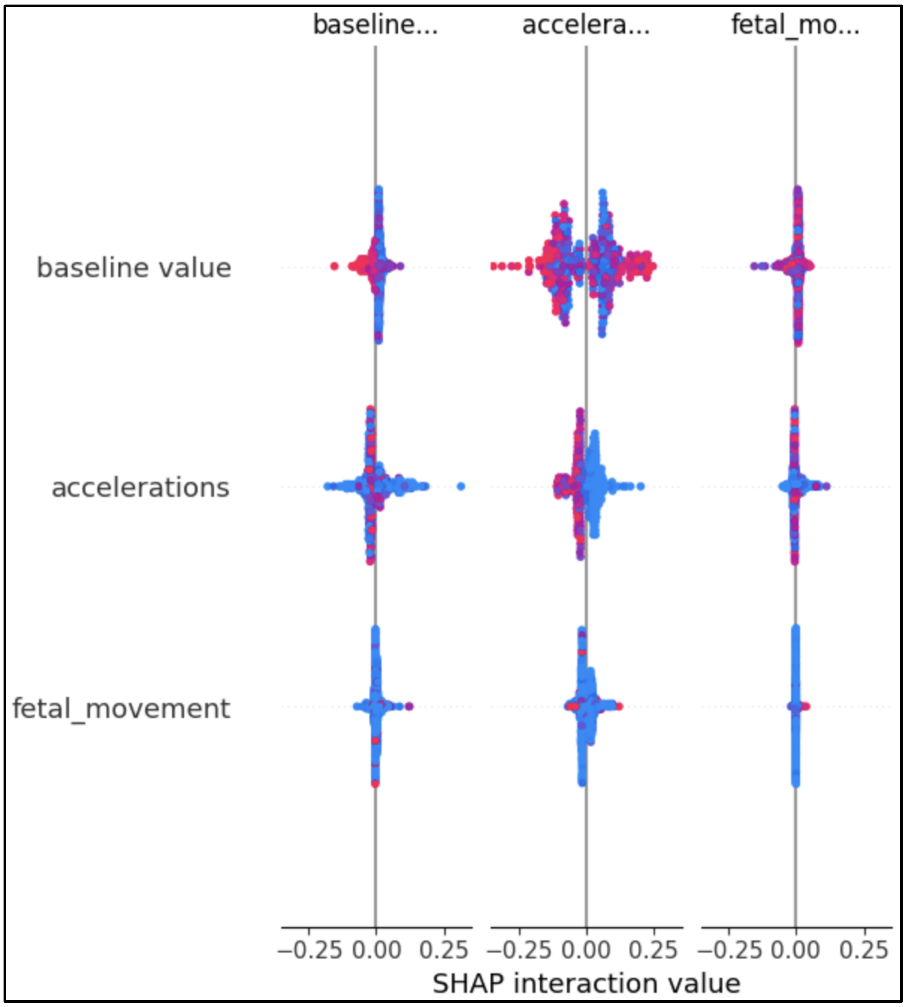
Beeswarm plot

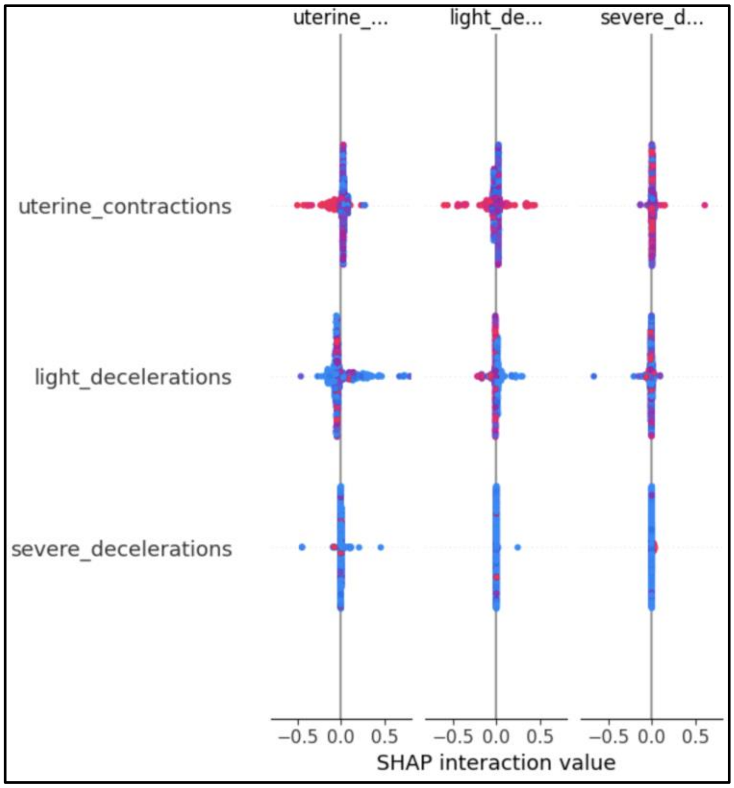


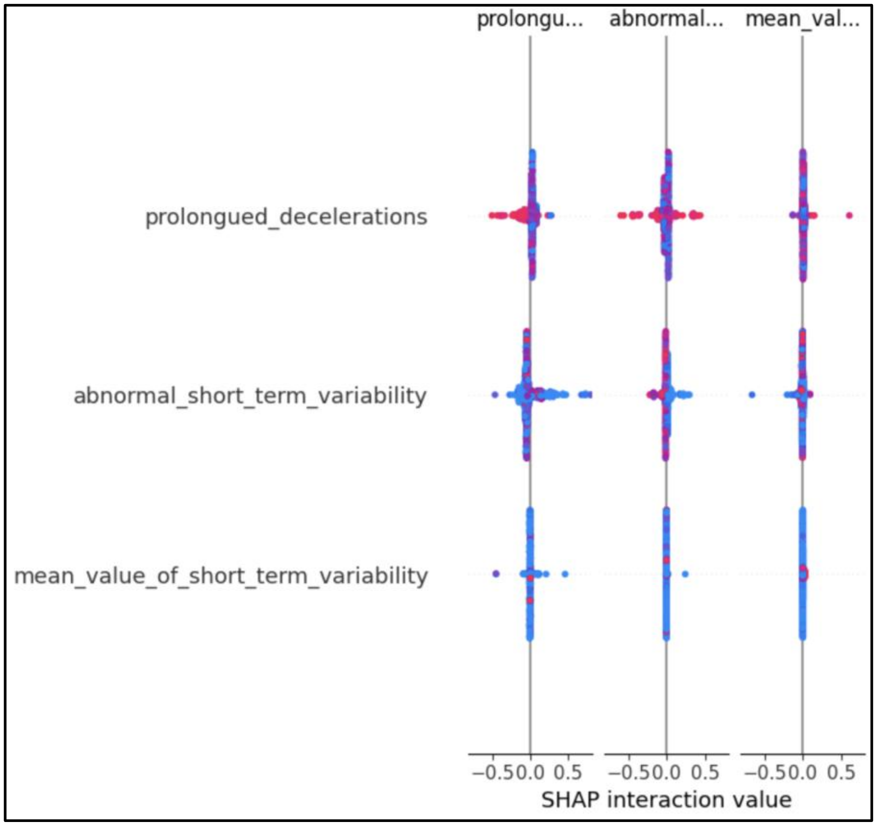
.



Summary Plot




uterine_...
light_de...
severe_d...
uterine_contractions
light_decelerations
severe_decelerations
−0.5 0.0 0.5
−0.5 0.0 0.5
−0.5 0.0 0.5
SHAP interaction value


prolongu...
abnormal...
mean_val...
prolongued_decelerations
abnormal_short_term_variability
mean_value_of_short_term_variability
−0.50.0 0.5
−0.50.0 0.5
−0.50.0 0.5
SHAP interaction value

FUTURE WORKS & CONCLUSION

Fashion MNIST

The future trajectories of this project involve leveraging the developed classifier and associated technologies to enhance various aspects of the fashion industry. One direction includes exploring advanced image analysis techniques like object detection and instance segmentation to provide nuanced insights into clothing items, improving the classification process. Another avenue is developing algorithms for personalized clothing recommendations based on user preferences and body measurements, elevating customer satisfaction and sales. Implementing virtual try-on experiences using AR or VR technologies can enrich the online shopping experience and reduce return rates. Additionally, investigating automated design generation with GANs could streamline the design process and foster innovation in fashion. The practical benefits of this model span from enhancing classification precision to driving AI-driven innovations in fashion, including generative design and user-centric virtual experiences.

In conclusion, this project conducted a thorough analysis of the Fashion-MNIST dataset, developed and evaluated a CNN image classifier, and explored its application in the fashion industry. The classifier's performance, highlighted using Grad-CAM, demonstrated its commendable accuracy of 91.2% on the test set. However, limitations were acknowledged, such as dataset complexity and resolution constraints. Future research should focus on addressing these limitations by applying the model to diverse datasets, enhancing interpretability, and mitigating class imbalances. These efforts aim to align the model's capabilities with the intricate requirements of the fashion industry.

Fetal Health

Moving forward, future research directions in fetal health assessment using CTGs include expanding the dataset, extracting data from CTG graphs, and exploring alternative algorithms to address limitations in data availability and model interpretability. Additionally, integrating additional features or data sources, such as maternal health indicators, and exploring advanced machine learning techniques like deep learning algorithms could enhance the model's predictive capabilities. Developing decision support systems and mobile health applications based on the validated model, along with exploring explainable AI techniques beyond SHAP values, are promising avenues for improving fetal health monitoring and prenatal care practices.

In conclusion, our study demonstrates the potential of predictive modeling and explainable AI techniques in enhancing fetal health assessment. By addressing challenges and exploring future research directions, we can further improve prenatal care and contribute to reducing child and maternal mortality rates globally. Efforts focused on expanding datasets, exploring advanced algorithms, and integrating additional features hold promise for advancing fetal health monitoring and improving outcomes for both mothers and infants.

ACKNOWLEDGEMENTS & REFERENCES

Acknowledgments

We would like to extend our gratitude to Dr. B Radhika Selvamani for entrusting us with the project on Explainable AI. Additionally, we express our gratitude to the authors and contributors of the referenced papers, whose work has served as a valuable resource for our research efforts.

References

1. Das, S. "Recurrent Neural Network Based Classification of Fetal Heart Rate Using Cardiotocography."

2. Dixit, R. R. "Predicting Fetal Health using Cardiotocograms: A Machine Learning Approach."

3. Miao, J. H., and Miao, K. H. "Cardiotocographic Diagnosis of Fetal Health based on Multiclass Morphologic Pattern Predictions using Deep Learning Classification."

4. Li, J. "Fetal Health Classification Based on Machine Learning."

5. Zhao, Z., et al. "DeepFHR: intelligent prediction of fetal Acidemia using fetal heart rate signals based on convolutional neural network."

6. Sahin, H., and Subasi, A. "Classification of Fetal State from the Cardiotocogram Recordings using ANN and Simple Logistic."

7. Tang, J., et al. "A Deep-Learning-Based Method Can Detect Both Common and Rare Genetic Disorders in Fetal Ultrasound."