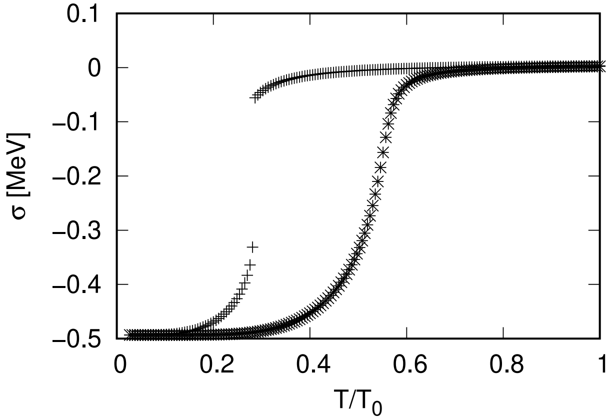
Hybrid Regression and Explainable AI for Phase Transition Analysis of two-flavour Quark Matter

Pradipta K. Banerjee[1], Tamal K. Mukherjee[2]

[1]Dept. of CSE-AIML, Future Institute of Technology
[2]Dept. of Physics, ADAMAS University

Objectives

- Identify the order of the phase transition (Crossover and 1st order)
- Classify the confined and deconfined phases and identify the T_C.
- Prediction of the CEP in the phase diagram (μ–T plane) of finite quark mass.



Multi Task Learning Methodology

1. Feature analysis using Random Forest and Explainable AI
2. Unsupervised clustering (GMM) of Crossover and First Order data
3. Phase classification using MLP(classification)
4. Transition temperature prediction using MLP (parametric regression)
5. Phase boundary prediction using KRR(semi-parametric regression)
6. Second order transition zone detection from soft probability assignment and predicted phase boundary

Multi Task Learning (MTL) Model

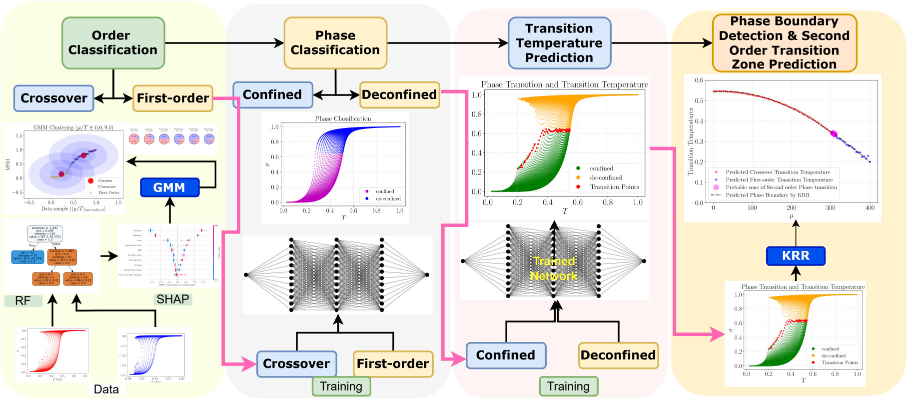


Figure: Multi task learning pipeline.

Feature Extraction and Importance Analysis using Explainable AI

- Proposed new feature Maximum Separation Measure(MSM) alongwith statistical and derivative features extracted from raw data
- The feature data set is divided into train and validation (train-valid split 80%-20%)
- Training of Random Forest and tested on validation set
- Feature importance analysis using SHAP based Explainable AI
- Selection of top-3 features for further clustering of Crossover and First-order

MSM

Given a signal vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, where x_i represents the i-th point in the sample:
If $\mathbf{x} \in \mathbb{R}^n$, the **Maximum Separability Measure (MSM)** is evaluated as:

$$\text{MSM}(\mathbf{x}) = \max_{i \in \{1, \ldots, n-1\}} \|x_{i+1} - x_i\|^2$$

This metric identifies the largest squared distance between consecutive points in the signal vector.

Random Forest and SHAP analysis of features

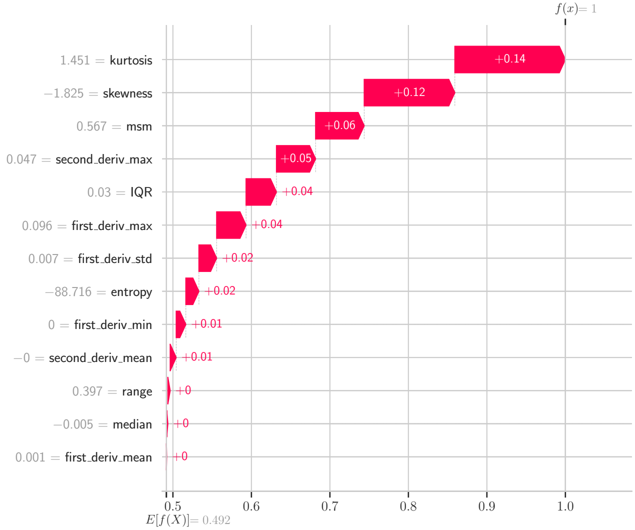


Figure: SHAP waterfall plot shows the importance of top-3 features towards RF prediction of single sample towards class-1(first-order).

Random Forest and SHAP analysis of features

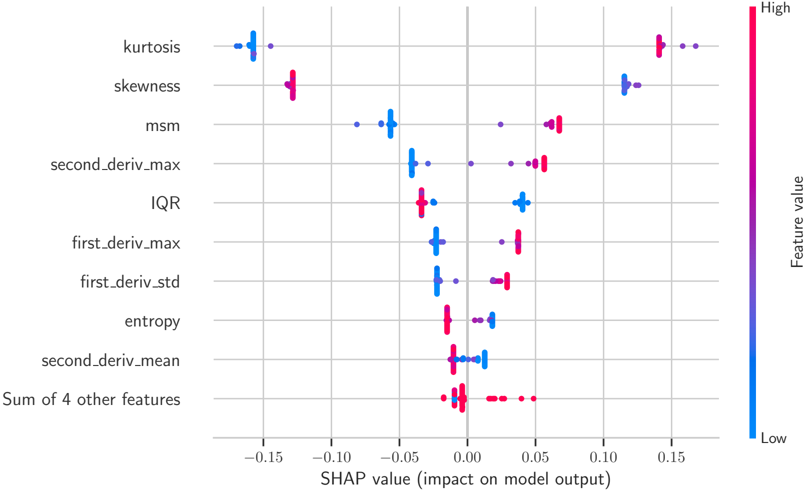


Figure: SHAP waterfall plot shows the importance of top-3 features towards RF prediction for all samples for binary classification problem.

Unsupervised Learning for clustering Crossover and First-order

SHAP based features

Main features selected:

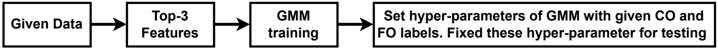
- Kurtosis
- Skewness
- Maximum Separation Measure(MSM)

SHAP is model-agnostic. Hence we choose top-3 features according to SHAP importance analysis for further study. Here we empolyed Gaussian Mixture Model.

Why GMM

- Can handle small data
- Can handle unlabeled data.
- Provides soft clustering by assigning probabilities to data points for belonging to clusters.

GMM based Clustering. *GMM working principle*



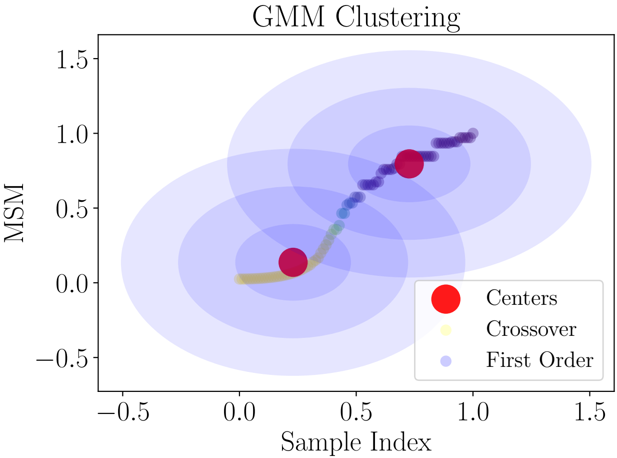


Figure: GMM based soft clustering of crossover and first-order data

Soft probability assignments by GMM

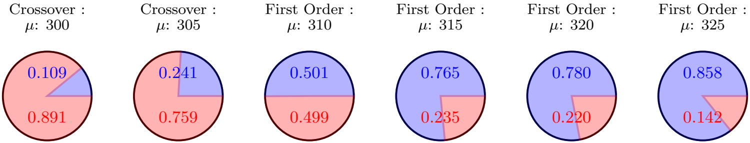


Figure: Soft probability assignment by GMM. It shows the probability assignment for classification crossover and first-order in the region of transition.

Soft-Class Assignment at Transition Region

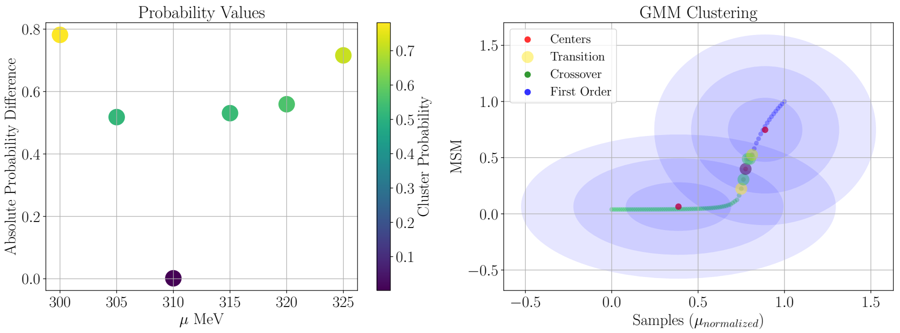


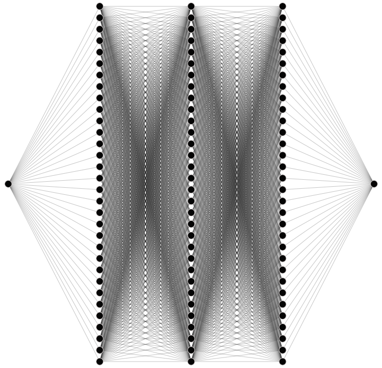
Figure: Left panel shows the low probability differences evaluated against soft class assignment for each sample at the transition zone. For each sample GMM assigns two probabilities (one for class-0 (crossover) and one for class-1(first-order). Right panel shows the transition zone and low confidence class assignments. Right panel shows that the transition occurs at the range $\mu = 300MeV$ to $\mu = 325$ MeV. Lowest probability difference obtained at $\mu = 310MeV$. From GMM model the transition from crossover to first-order occurs at $\mu = 305MeV$.

Hybrid Regression Technique for Phase Boundary prediction

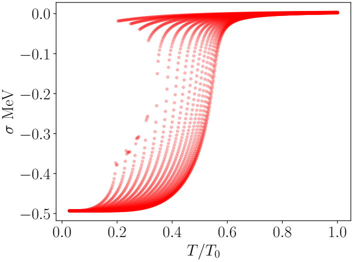
- Both parametric and semi-parametric method is used to find the phase boundary by means of finding transition temperatures.
- Once classified as crossover or first-order, the proposed MTL model use Multi layer Perceptron to classify the confined and de-confined phase.
- For a fixed coupling constant, from a set of $\sigma - T$ curves the transition temperatures are predicted using MLP. This is parametric regression process.
- Once transition temperature obtained a semi-parametric method, Kernel ridge Regression is exploited to obtain the phase boundary.

MLP Model Architecture. *Detail MLP training block diagram*

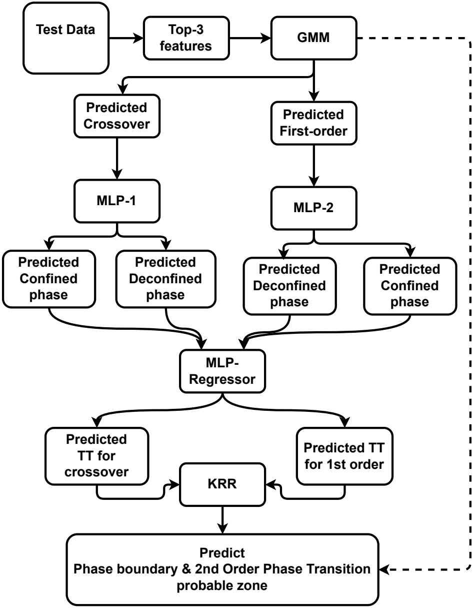
MLP model, input $\in \mathcal{R}$, output $\in 0, 1$, Model shows only one sample having dimension $\mathcal{R}$ (Batch-wise training is done. Batch size=64), activations = ReLU, output activation = sigmoid (as binary classification problem) , hidden size=32 (empirically set). Same architecture used for Phase classification (MLP as classifier) and Transition temperature prediction (MLP as regressor). The output dimension of the MLP regressor $\in \mathcal{R}$.



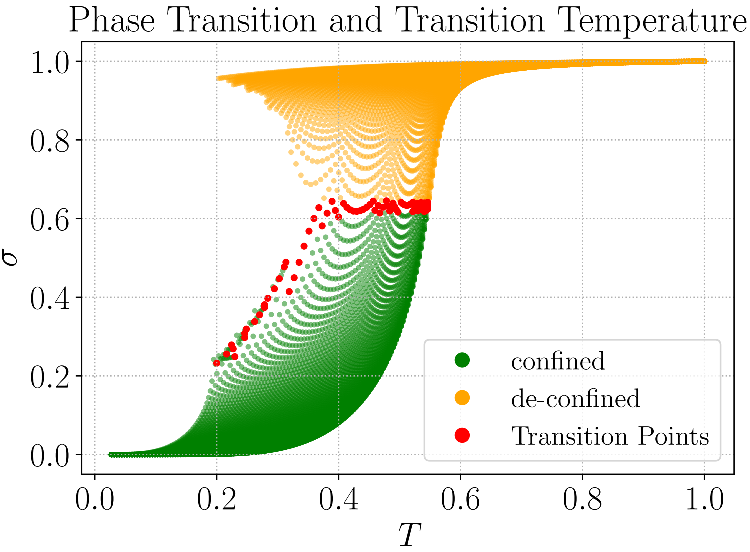
Training Data



Experiment and Results: Test phase block diagram



Experimental Results: Phase classification and Transition Temperature Prediction



Experimental Results: Phase boundary prediction: KRR. *Soft probability assignments*

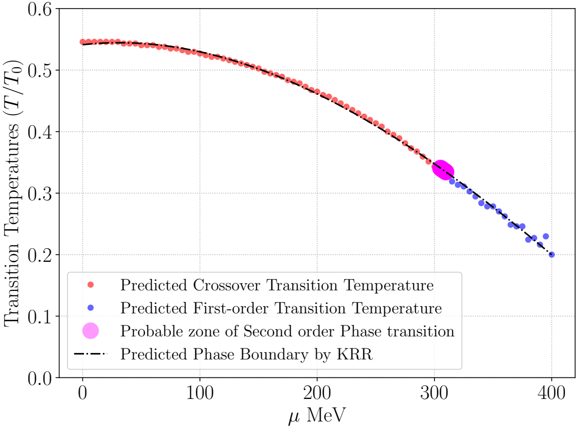


Figure: Kernel Ridge Regression (KRR) is used to fit the phase boundary with the help of predicted transition temperatures (obtained by MLP regressor) for the crossover and first-order data. Probable Zone for 2nd order phase Transition is reported ; depending on soft probability assignment by GMM

Experimental Results: Comparative study in phase boundary prediction

Classification result for confined and deconfined phase using trained MLP. Prediction results for phase boundary and second order transition zone obtained by KRR. σ scaled by min-max scaling method and $T_0 = 0.37 GeV$. Silhoutte Score of GMM clustering for the corresponding dataset is obtained as $\mathbf{0.7215}$ at and average error rate (ARR) for transition temperature prediction is $\mathbf{0.64\%}$.

Table: Regression Methods and Mean Squared Errors. Comparative study of different regression techniques for phase boundary predictions. KRR gives the lowest MSE.

Method	**MSE**
Polynomial Regression	0.162551
Kernel Ridge Regression	$\mathbf{0.000004}$
KNN Regression	0.000091
Hybrid Model	0.000033
Gaussian Process Regression	0.000140

References

G. Nambu, Y.; Jona-Lasinio.
Dynamical model of elementary particles based on an analogy with superconductivity. i ii.
Physical Review, 122 124(1):345–358 246–254, 1961.

M. Buballa.
Njl-model analysis of dense quark matter.
Physics Reports, 407:205–376, 2005.

Wenjia Wang and Bing-Yi Jing.
Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression.
Journal of Machine Learning Research, 23(193):1–67, 2022.

Yi Zhang, Miaomiao Li, Siwei Wang, Sisi Dai, Lei Luo, En Zhu, Huiying Xu, Xinzhong Zhu, Chaoyun Yao, and Haoran Zhou.
Gaussian mixture model clustering with incomplete data.
ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17(1s):1–14, 2021.

Thank you

Thank You!

Supplementary

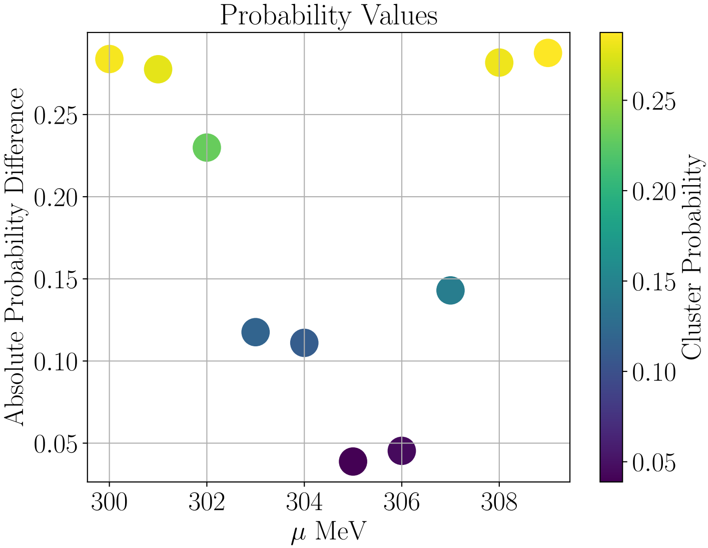


Figure: Low confidence in terms of absolute probability differences at transition region.

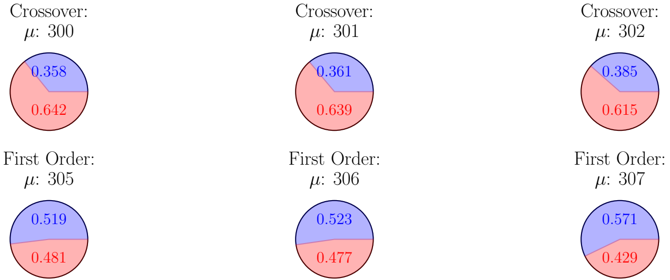
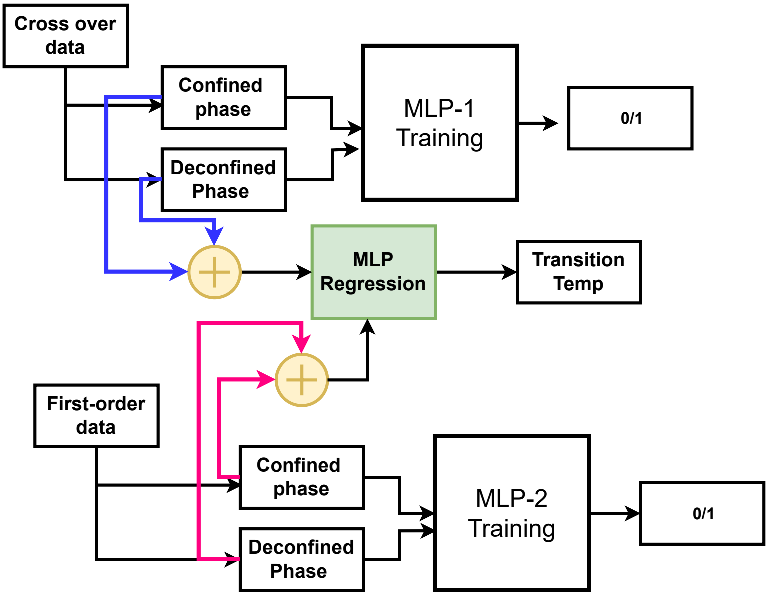


Figure: Low confidence in terms of absolute probability differences at transition region. ABS probability difference <0.3

MLP Training Process. *MLP model*



Gaussian Mixture Model (GMM) Training Block Diagram. *GMM based clustering*

- **Initialization:**
  - Initialize number of components (K). Set initial parameters: means (μ_k), covariances (Σ_k), and weights (π_k).
- **Expectation Maximization Algorithm:**
  - **E-Step:** Compute responsibilities:

$$\gamma_{ik} = \frac{\pi_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$

  - **M-Step:** Update parameters:

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}}, \quad \Sigma_k = \frac{\sum_{i=1}^{N} \gamma_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} \gamma_{ik}},$$

$$\pi_k = \frac{\sum_{i=1}^{N} \gamma_{ik}}{N}$$

- **Convergence Check:** Stop when log-likelihood improves below a threshold.
- **Output:** Final parameters: (μ_k), (Σ_k), (π_k).

Definitions of various features used in MTL

Features	Mathematical Expression
MSM	$\text{MSM} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$
Median	$\text{Median} = \text{Median}(\{x_1, x_2, \ldots, x_n\})$
Range	$\text{Range} = \max(x) - \min(x)$
IQR	$\text{IQR} = Q_3 - Q_1$
Skewness	$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}}$
Kurtosis	$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{2}}$
Entropy	$\text{Entropy} = -\sum_{i=1}^{k} p_i \log(p_i)$
First Derivative Mean	$\text{Mean}(\Delta x) = \frac{1}{n-1}\sum_{i=1}^{n-1}(x_{i+1} - x_i)$
First Derivative Std	$\text{Std}(\Delta x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}\left[(x_{i+1} - x_i) - \text{Mean}(\Delta x)\right]^2}$
First Derivative Max	$\max(\Delta x)$
First Derivative Min	$\min(\Delta x)$
Second Derivative Mean	$\text{Mean}(\Delta^2 x) = \frac{1}{n-2}\sum_{i=1}^{n-2}(x_{i+2} - 2x_{i+1} + x_i)$
Second Derivative Max	$\max(\Delta^2 x)$

Table 1: Features used for Random Forest Classifier and Explainable AI