LIP-STUDENTS-25-27

Investigating the Flavour Anomalies with Machine Learning at LHC

Gabriela Sousa[1,a] and Teresa Escária[2,b]

[1] *Instituto Superior Técnico*
[2] *Imperial College London*

Project supervisor: Alessio Boletti and Nuno Leonardo

November 21, 2025

Abstract. While the Standard Model (SM) accurately describes most particle interactions, certain rare B-meson decays show deviations, known as flavour anomalies, particularly in $b \to s\ell^+\ell^-$ transitions. This study investigates the resonant decay channel $B^0 \to K^{*0}J/\Psi$ using CMS Run-3 data. Candidates are reconstructed from two muons and a $K^+\pi^-$ pair, and a binned likelihood fit is performed to extract signal and background yields. To enhance discrimination, single-variable selections and a multivariate analysis with a neural network (NN) trained on 13 optimised features are applied. The NN selection nearly doubles the S/B ratio and increases the statistical significance by ∼30%. The framework developed here can be directly applied to the non-resonant decay $B^0 \to K^{*0}\mu^+\mu^-$ to explore potential new physics.

Keywords: CMS, Heavy-flavour Physics, Neural Network, New Physics, LFU

1 Introduction

1.1 The Flavour Anomalies ($b \to s\ell^+\ell^-$)

The study of rare decays has long provided crucial insights into the structure of the Standard Model (SM) and potential effects of new physics. In particular, processes involving flavour-changing neutral currents are of special interest, since they are forbidden at tree level in the SM and occur only through higher-order electroweak loop diagrams, like penguin or box diagrams.

Within this class, transitions like $b \to s\ell^+\ell^-$ play a central role. Their branching fractions are highly suppressed, yet they can still be predicted with relatively small uncertainties, which makes them excellent probes of physics beyond the SM.

Several extensions of the SM predict new heavy mediators that could virtually contribute to these processes. Examples include a heavy neutral gauge boson (Z') or a leptoquark (LQ). Such particles would modify the effective couplings of the $b \to s\ell^+\ell^-$ transition, potentially altering the overall decay rates, distorting the angular distributions of the final-state particles, or breaking lepton flavour universality (LFU) by interacting differently with different lepton species.

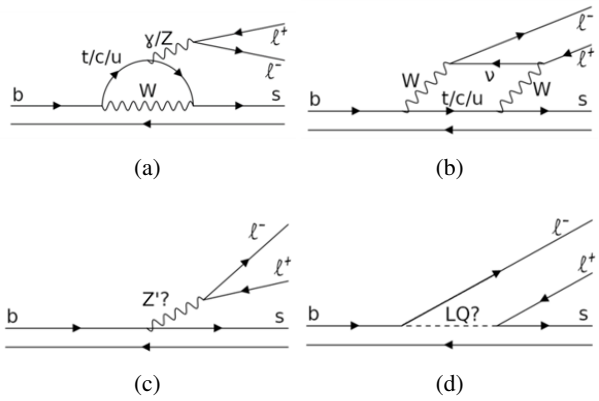


Figure 1: Illustration of SM contributions (top) through box and penguin diagrams (Fig. 1b and Fig. 1a, respectively), as well as possible New Physics contributions to $b \to s\ell^+\ell^-$ transitions, such as a Z' boson (Fig. 1c) or a leptoquark (Fig. 1d).

1.2 Angular Analysis

Beyond the study of branching fractions, rare $b \to s\ell^+\ell^-$ are particularly powerful because they allow for detailed angular analysis of the final-state particles. By constructing suitable observables from angular distributions, it is possible to separate short-distance dynamics from hadronic effects, thus maximising sensitivity to contributions from physics beyond the SM.

One of the most prominent examples is the observable P'_5, which is designed to reduce theoretical uncertainties while retaining a strong sensitivity to possible new interactions. Measurements of P'_5 in the decay $B^0 \to K^{*0}\mu^+\mu^-$ have shown deviations from the SM predictions in specific dilepton mass regions. These discrepancies, first reported by LHCb [1, 2] and more recently confirmed by CMS using Run-2 data [3], are commonly referred to as flavour anomalies.

Although the statistical significance of these anomalies is not yet strong enough to claim a discovery, their consistency across various experiments and datasets makes them one of the most promising hints of new physics in the flavour sector. This motivates further studies using larger and more precise datasets, such as the Run-3 data from the LHC.

[a] e-mail: gabriela.v.sousa@tecnico.ulisboa.pt
[b] e-mail: teresa.escaria23@imperial.ac.uk

2 The $B^0 \to K^{*0}\mu^+\mu^-$ decay

The decay $B^0 \to K^{*0}\mu^+\mu^-$ can proceed through two different mechanisms that lead to the same final state. The first is the non-resonant channel, with the following decay:

$$B^0 \to K^{*0}\mu^+\mu^- \to K^+\pi^-\mu^+\mu^-.$$

This transition is loop suppressed in the SM, therefore, it is highly sensitive to potential contributions from new physics like the ones mentioned in Section 1.1. Angular analyses, like the measurement of the observable P'_5, are performed in this channel to study deviations from the SM.

The second mechanism is the resonant channel:

$$B^0 \to K^{*0}J/\psi \to K^+\pi^-\mu^+\mu^-.$$

This process occurs already at tree level in the SM, so it is very well understood. Most importantly, its branching fraction has been precisely measured [4], so the probability that this transition will occur is known. For this reason, this decay serves as a control channel because it provides a clean reference sample that can be used to validate the reconstruction and selection features before extending the analysis to the more sensitive non-resonant channel.

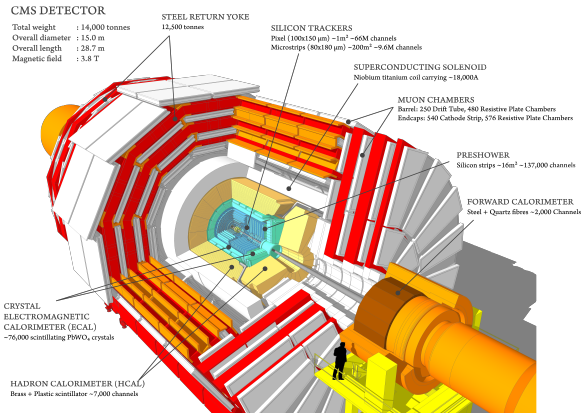
The present work focusses on the resonant channel to establish a robust control sample that will support future studies of the non-resonant decay.

3 The CMS Detector

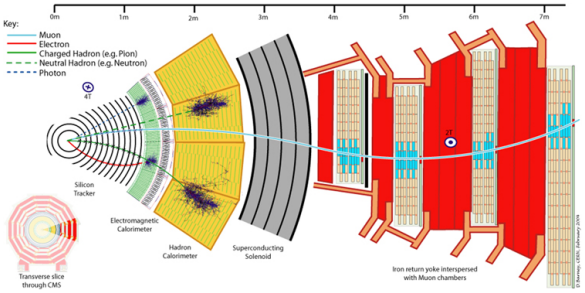
The Compact Muon Solenoid (CMS), represented in Fig. 2, is a general-purpose detector at the Large Hadron Collider (LHC). Its geometry is cylindrical, with layered sub-detectors arranged concentrically around the interaction point [5].

Closest to the beamspot, i.e. the central region where the collision between the proton beams occurs, lies the silicon tracker, composed of pixel and strip detectors, which reconstructs charged-particle trajectories. The curvature of the measured tracks in the magnetic field allows for momentum and charge determination, as, in the same magnetic field, positively and negatively charged particles bend in opposite directions, and high-momentum particles bend less compared to low-momentum ones. Surrounding the tracker, the electromagnetic calorimeter (ECAL) measures the energies of electrons and photons via electromagnetic showers, whereas the hadronic calorimeter (HCAL) is dedicated to capturing the energy deposited by hadrons. Continuing outwards, there is the superconducting solenoid, which generates a uniform 3.8 T magnetic field, essential for the particle trajectory analysis. Lastly, located in the outermost part of the detector, the muon chambers, consisting of up to four gas-ionisation stations embedded in the steel return yoke, identify and measure muons, which traverse the calorimeters with minimal energy loss.

For the purpose of this study of the $B^0 \to K^{*0}\mu^+\mu^-$ decay, the two key subsystems are the silicon tracker (responsible for charged particle trajectory reconstruction, determine their momenta, and reconstruct vertices for the desired decay), and the muon chambers (which trigger and enable a full muon reconstruction).



(a) CMS detector schematic



(b) Transverse view of CMS

Figure 2: Schematic views of the CMS detector.

4 Reconstruction and Selection

Using the tracking and muon systems of CMS (Section 3), the $B^0 \to K^+\pi^-\mu^+\mu^-$ candidates are reconstructed as follows. Two oppositely charged muons are combined with two oppositely charged tracks and a common displaced vertex fit is performed with these four tracks. The hadron pair is assigned kaon and pion mass hypotheses and combined to form a K^{*0} candidate, which is then combined with the dimuon system to build a B^0 candidate. Standard requirements on track quality, particle identification, and vertex fit probability are applied.

Since both $B^0 \to K^+\pi^-\mu^+\mu^-$ and its charge conjugate decay, $\bar{B}^0 \to K^-\pi^+\mu^+\mu^-$, can contribute to the same reconstructed candidate, each pair of hadron tracks has exactly two possible mass assignments. A particular challenge in CMS is the absence of a dedicated hadron particle identification, which prevents a direct distinction between the kaon and pion tracks. To resolve this, the assignment is chosen such that the invariant mass of the hadron pair is closest to the known K^{*0} mass ($m_{PDG}(K^*(892)) = (6.7 \pm 1.2)$ MeV [4]). This procedure is effective, but a small fraction of candidates are mis-tagged, resulting in a residual contribution from incorrectly assigned tracks.

The relatively long lifetime of the B^0 meson allows its decay to be distinguished by a secondary vertex, displaced from the primary interaction point. This property provides a strong handle for identifying true B^0 decays.

In spite of these selections, background arises from two main sources: combinatorial background, which results from random combinations of tracks that mimic a B^0 candidate, and physics background, originating from other real B-meson decays that have similar final states.

5 Data Preparation

A sample of proton-proton collisions collected by the CMS experiment, during Run-3 (2022), as well as a Monte Carlo realistic simulation of the resonant decay are analysed, in order to extract the signal and background yields for the $B^0 \to K^{*0} J/\Psi$ channel.

Initial inspection of the B^0 mass spectrum in the data reveals no clear signal peak and a background-dominated distribution, as shown in Fig. 3.

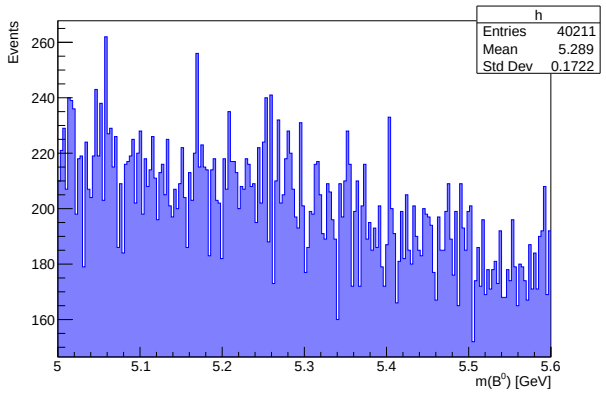


Figure 3: B^0 mass distribution in the data sample before any selection is applied.

5.1 MC Binned Likelihood Fit

To model the reconstructed signal in simulation, a binned maximum likelihood fit is performed to the B^0 mass distribution in Monte Carlo. Separate probability density functions (PDFs) are used for the right-tagged (RT) and wrong-tagged (WT) signal candidates, which arise from the kaon-pion assignment procedure described in Section 4.

Two independent fits are performed, one to the RT sample and one to the WT sample, each using its corresponding PDF model. For each component, the mass shape is parametrised as the sum of a Gaussian distribution, f_G, and a Crystal Ball (CB) function, f_{CB}, with a common mean:

$$\mathcal{P} = C f_{CB} + (1 - C) f_G, \tag{1}$$

where C is the fraction of the CB contribution.

The Gaussian PDF is given by

$$f_G(m; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(m-\mu)^2}{2\sigma^2}\right), \tag{2}$$

and the Crystal Ball function is defined by a Gaussian core and a power-law low-end tail:

$$f_{CB}(m; \mu, \sigma, \alpha, n) = \mathcal{N}_{CB} \begin{cases} \exp\left(-\frac{(m-\mu)^2}{2\sigma^2}\right), & \frac{m-\mu}{\sigma} > -\alpha, \\ A\left(B - \frac{m-\mu}{\sigma}\right)^{-n}, & \frac{m-\mu}{\sigma} \leq -\alpha, \end{cases} \tag{3}$$

where $\mathcal{N}_{CB}$ is a normalisation factor and A and B are constants determined from α and n.

All fits are implemented using the RooFit toolkit from the ROOT framework (CERN) [6]. For the initial RT and WT fits, the CB fraction C and the widths of the Gaussian and CB components are allowed to float for each category. Once the RT and WT shapes are determined, these parameters are fixed to define the final signal-shape used in the analysis and illustrated in Fig. 4.

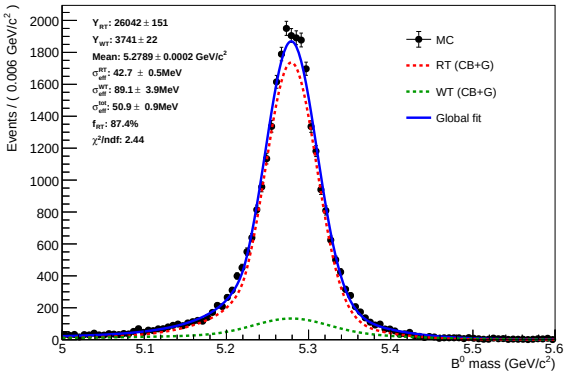


Figure 4: Fitted RT and WT MC mass distributions with the corresponding CB+Gaussian PDFs.

In the RT fit, the CB fraction is found to be $C = (0.839 \pm 0.007)$, corresponding to 83.9% Crystal Ball and 16.1% Gaussian distribution. The WT component yields $C = (0.533 \pm 0.036)$, i.e. 53.3% CB and 46.7% Gaussian. From the combined signal fit, an RT event fraction of 87.4% is obtained.

5.2 Relevant Variables

After obtaining a well-defined signal shape in simulation, the next step is to extract the signal in the real data sample, which is dominated by background. To enhance the signal-to-background ratio, several variables are used in the analysis (listed below). These variables, along with a schematic representation in Fig. 5, illustrate the relevant features of the decay and help to better understand the underlying physics.

- Flight Length: distance between the primary vertex and the secondary vertex (B^0 decay vertex);

- $\cos\alpha$: cosine of the angle between the reconstructed B^0 meson momentum and its flight direction;
- Vertex confidence level: likelihood that the four tracks originate from a common vertex;
- Distance of Closest Approach (DCA) variables: how close a track comes to the beamline (Fig. 5 illustrates the shortest distance from the beamspot (BS) to the positive hadron trajectory, d_0);
- θ : polar angle of the particle trajectory;
- $\eta = -\ln(\tan\theta/2)$: pseudorapidity;
- $\phi = \tan^{-1}\left(\frac{p_y}{p_x}\right)$: azimuthal angle;
- $p_T = \sqrt{p_x^2 + p_y^2}$: transverse momentum;
- $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$: separation between the trajectories of two particles;
- Isolation variables: quantify the amount of activity in a cone of $\Delta R = 0.4$ around the candidate particle (muon or hadron track);
  - Absolute isolation: scalar sum of the transverse momenta of all other tracks in the cone;
  - Relative isolation: ratio of the absolute isolation to the candidate transverse momentum;
  - Combined relative isolation: sum of the relative isolations of the four tracks forming the B^0 candidate. This variable characterizes the overall isolation of the full B^0 decay.

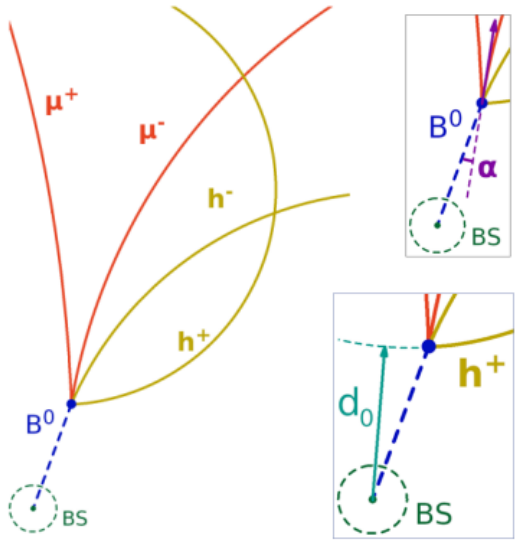


Figure 5: Schematic view of the $B^0 \to K^{*0} J/\Psi$ decay.

5.3 Signal and Background Samples

To assess which variables are most discriminating, high-purity samples of signal and background are required. The signal sample is taken from simulation, whereas the background is estimated from data using regions adjacent to the signal window, known as sidebands. The sideband limits are defined using an effective standard deviation obtained from the fits described in Section 5.1. For each RT and WT component, the standard deviation is determined from the Gaussian and CB widths (σ_G and σ_{CB}), weighted by the CB fraction, C:

$$\sigma_{R(W)T} = \sqrt{C\sigma_{CB}^2 + (1-C)\sigma_G^2}. \quad (4)$$

The effective standard deviation of the combined signal model is then obtained similarly, using the RT and WT standard deviations (σ_{RT} and σ_{WT}), and the RT event fraction, f_{RT}:

$$\sigma_{\text{eff}} = \sqrt{f_{RT}\sigma_{RT}^2 + (1-f_{RT})\sigma_{WT}^2}. \quad (5)$$

From the MC fit shown in Fig. 4, a total effective width of $\sigma_{\text{eff}} = (50.9 \pm 0.9)$ MeV is obtained. The signal region is defined as $\pm 3\sigma_{\text{eff}}$ around the mean, while the left and right sidebands are taken as $m \in [5, 5.126]$ GeV and $m \in [5.432, 5.6]$ GeV, respectively. The background sample is thus composed of the sidebands in data, whereas the signal sample corresponds to the simulated dataset, restricted to the region between the sidebands, as illustrated in Fig. 6.

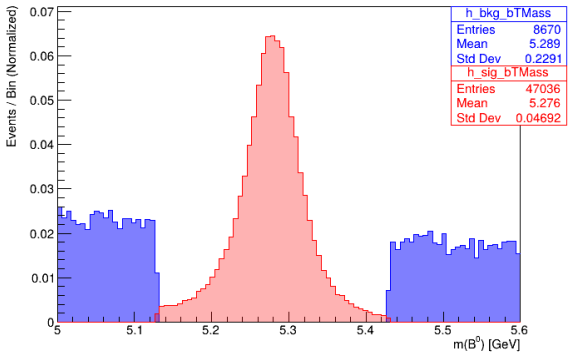


Figure 6: Normalised mass distributions for the MC signal sample (red) and the background sample from data sidebands (blue).

5.4 Variable distributions

The distributions of signal and background are compared across the relevant variables. Based on visual inspection, simple selection criteria are applied to suppress background. In particular, thresholds are placed on the vertex confidence level (bVtxCL $>$ 0.20) and the flight length (bLBS $>$ 0.05), as shown in Fig. 7.

5.5 Data Binned Likelihood Fit

After the selection criteria described in the previous section are applied, a clear signal peak becomes visible in the data B^0 mass distribution. A binned likelihood fit is then performed. The signal PDF is constructed by fixing the

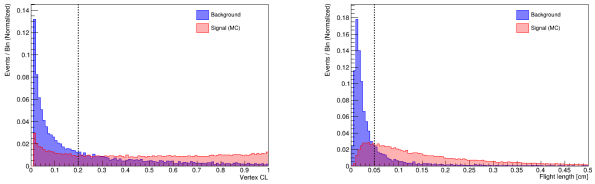


Figure 7: Normalised distributions of the vertex confidence level (left) and flight length (right), with the applied selection thresholds indicated.

RT and WT shapes, as well as the RT event fraction, f_{RT}, as determined in Section 5.1:

$$\mathcal{P}_S(m) = f_{RT}\,\mathcal{P}_{RT}(m) + (1 - f_{RT})\,\mathcal{P}_{WT}(m), \qquad (6)$$

where $\mathcal{P}_{RT}$ and $\mathcal{P}_{WT}$ are defined by (1).

The background component is given by an exponential function:

$$\mathcal{P}_B \equiv f(m;\lambda) = \mathcal{N}\exp(\lambda m)\,, \qquad (7)$$

with slope parameter $\lambda < 0$ and normalization $\mathcal{N}$.

The total PDF is then expressed as a weighted sum of signal and background:

$$\mathcal{P}_G(m) = Y_S\,\mathcal{P}_S(m) + Y_B\,\mathcal{P}_B(m), \qquad (8)$$

where Y_S and Y_B denote the signal and background yields, respectively.

The fit, shown in Fig. 8, returns $Y_S = (595 \pm 38)$ and $Y_B = (1871 \pm 52)$ events. The fitted mean is $\mu = (5278.93 \pm 0.24)$ MeV, in good agreement with the B^0 mass reported by the Particle Data Group (PDG), $m = (5279.63 \pm 0.20)$ MeV [4].

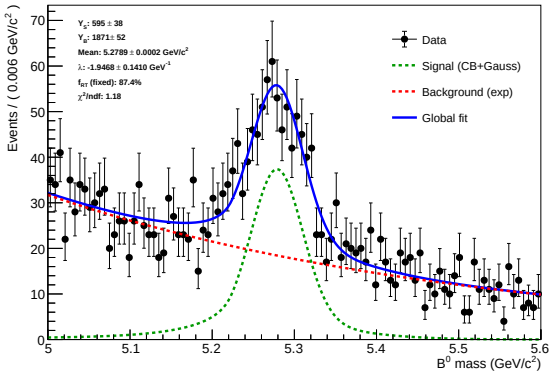


Figure 8: Binned likelihood fit to the B^0 mass distribution in CMS Run-3 data, showing the signal and background components.

These fitted yields and parametrisations serve as the baseline for validating the multivariate analysis strategy, where additional discriminating power is obtained through a neural network classifier.

6 NN Multivariate Analysis

6.1 Feature Selection

Before training the neural network classifier, it is important to carefully select the input variables. Including too many features, especially those carrying redundant or highly correlated information, can reduce the performance of the model and increase the risk of overfitting.

The goal of feature selection is to optimise the set of input variables by keeping those that provide independent discriminating power and eliminating redundant ones. To this end, two complementary methods are applied: the study of correlations among the variables and the evaluation of their relative importance through SHAP values [7].

6.1.1 Correlation Matrix

The first step consists in computing the correlation matrix of all candidate input variables. This provides a quantitative measure of how strongly each pair of features is correlated, with values ranging from −1 (perfect negative correlation) to +1 (perfect positive correlation). Variables that are strongly correlated typically carry overlapping information and, therefore, including both in the training does not improve the classifier's performance.

The correlation matrix thus offers a straightforward way to characterise dependencies among features before proceeding with the selection process.

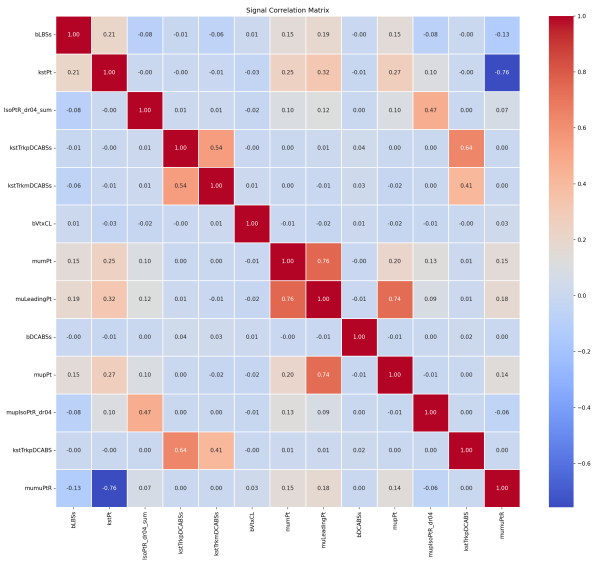


Figure 9: Signal correlation matrix of the final set of selected variables. The colour scale indicates the linear correlation coefficient between each pair of features, ranging from −1 (strong negative correlation, blue) to +1 (strong positive correlation, red).

6.1.2 Cumulative SHAP importance

Feature importance is combined with the correlation study to obtain a compact, non-redundant input set. Starting

from the full list of variables, those that are highly correlated ($|\rho| \geq 0.8$) are first grouped together using the correlation matrix. Within each correlated group, a single representative is chosen: the variable with the highest SHAP importance.

The ranked representatives are then used to construct the cumulative importance curve shown in Fig. 10. On the horizontal axis, the representatives are added one by one in order of decreasing SHAP score, while the vertical axis displays the fraction of total importance explained by the variables included up to that point.

The curve rises sharply at the beginning, indicating that only a few features already account for most of the performance, and then gradually flattens. The horizontal red dashed line corresponds to the point where the cumulative importance reaches 95% of the total, which is taken as the target coverage. The vertical green line indicates the minimum number of representatives required to reach this threshold. The selected variables correspond to those lying to the left of the green line, highlighted in the shaded region.

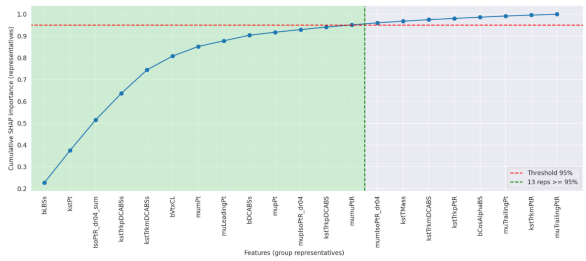


Figure 10: Cumulative SHAP importance for correlation-group representatives, sorted by SHAP score. The red dashed line marks the 95% target; the vertical green dashed line indicates the minimal number of variables needed to reach it. The selected features are those displayed on the x axis to the left of the green dashed line (green-shaded region).

6.2 Neural Network Classifier

To improve signal-to-background separation, a Neural Network Classifier is trained using PyTorch. The classifier takes as input the 13 features selected in Section 6.1.2 and processes them through three fully connected hidden layers with 45, 45 and 56 neurons respectively (Fig. 11). Each hidden layer is followed by a dropout layer, which randomly deactivates about 7% of its neurons during training, to mitigate overfitting and improve generalisation to unseen data. ReLU (Rectified Linear Unit) activations are applied in the hidden layers, allowing the network to learn non-linear patterns, while the output layer uses a sigmoid activation function (9), which maps the final score to a probability between 0 and 1, representing the likelihood of an event being signal.

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{9}$$

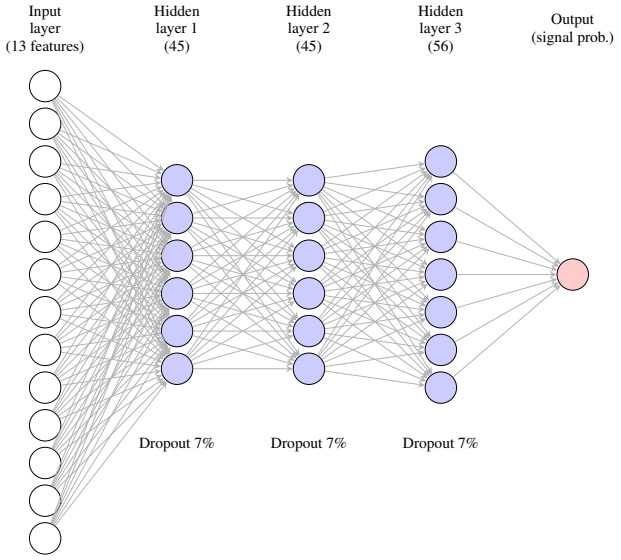


Figure 11: Schematic representation of the Neural Network Classifier: 13 input features, three hidden layers (45, 45, 56 neurons), and a single output node giving the signal probability. Only a subset of neurons per hidden layer is shown for clarity.

6.2.1 *Model Optimisation: Optuna*

The performance of such networks is highly sensitive to the choice of hyperparameters. To identify the optimal configuration, an automated optimisation is performed using the Optuna framework [8].

The hyperparameters explored include the number of layers, number of neurons per layer, dropout fraction, learning rate, weight decay, and batch size. The learning rate controls the size of the parameter updates during gradient descent, the weight decay introduces an additional penalty on large weights to regularise the network and reduce overfitting, and the batch size determines how many training examples are processed before the model's parameters are processed.

A total of 50 training trials are carried out, where these values are sampled within predefined ranges. The configuration that yields the lowest validation loss (i.e. the smallest error between the model predictions and the target values on a validation set) is chosen as the best model. The hyperparameter ranges and corresponding optimal values are summarised in Table 1.

Table 1: Hyperparameter ranges explored with Optuna and the corresponding optimal values selected for the neural network classifier.

Hyperparameter	**Range**	**Optimal value**
Number of hidden layers	1–5	3
Neurons per layer	8–128	45, 45, 56
Dropout fraction	0.0–0.5	0.066
Learning rate	10^{-5}–10^{-2}	8.9×10^{-4}
Weight decay	0–10^{-3}	1.9×10^{-4}
Batch size	{32, 64, 128, 256}	128

6.3 Training Procedure

The neural network is trained on the prepared dataset using the following strategy.

6.3.1 Optimiser

Training is performed with the Adam optimiser. The learning rate and weight decay values are taken from the Optuna optimisation.

6.3.2 Loss Function and Class Imbalance

To account for the imbalance between background and signal events, a weighted binary cross-entropy loss function is employed. Let $y \in \{0, 1\}$ denote the target label and $\hat{y}$ the predicted probability. The standard binary cross-entropy loss is given by

$$\ell(y, \hat{y}) = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]. \quad (10)$$

Class weights α_0 and α_1 are introduced, chosen to be inversely proportional to the number of training samples in each class ($\alpha_c \propto 1/n_c$), and normalised such that $\alpha_0 + \alpha_1 = 1$. The balanced loss becomes

$$\ell_{balanced}(y, \hat{y}) = -[\alpha_1 \, y \log(\hat{y}) + \alpha_0 \, (1-y) \log(1-\hat{y})]. \quad (11)$$

This reduces to (10) in the unweighted case and ensures that minority-class events are up-weighted, while majority-class events are down-weighted, thereby preventing the model from being biased toward the more prevalent class. The final training loss is computed as the mean of the per-sample balanced cross-entropy values.

6.3.3 Batch Size and Epochs

The dataset is divided into mini-batches of 128 events, and the training proceeds for up to 500 epochs. These values balance convergence stability and computational efficiency.

6.3.4 Validation and Early Stopping

At each epoch, a fraction of the training sample (about 25%) is reserved for validation. Early stopping measures are applied, such that if the validation loss has not improved for 100 consecutive epochs, training stops and the model's best state is restored. Fig. 12 shows the training and validation loss curves, monitored during training. The stable decrease indicates the model is learning properly without overfitting.

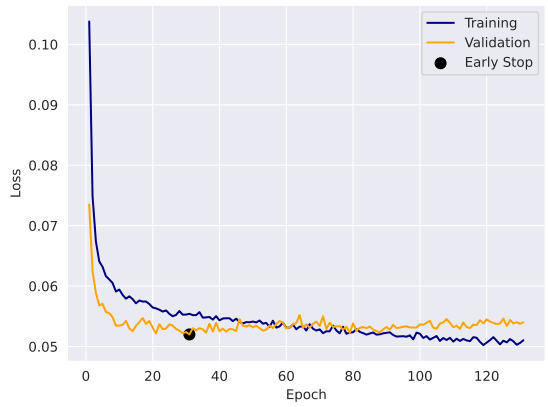


Figure 12: Training loss curve (blue): measures the error between the model's predictions and the target values in the training dataset (50% of full dataset); validation loss (yellow): measures the error on a separate set (validation set) that the model has not seen during training.

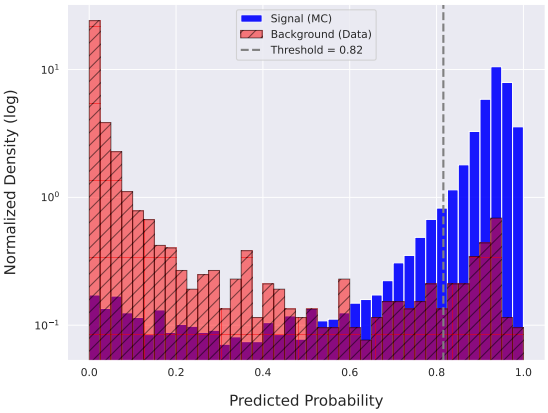


Figure 13: Distribution of NN output probabilities for signal and background samples. Signal peaks near 1, while background peaks near 0.

6.4 Neural Network Output

The NN output, i.e. the probability assigned to each event being signal, is illustrated in Fig. 13. Ideally, signal events should accumulate near 1, and background events cluster near 0. The observed distribution shows a clear separation, indicating that the classifier effectively captures the differences between signal and background.

The performance of the classifier is evaluated using Receiver Operating Characteristic (ROC) curves (Fig. 14). These plot signal efficiency, or true positive rate (TPR), against the background misclassification rate, or false positive rate (FPR). The curve is built by sweeping through all possible classifier thresholds: the closer it lies to the top left corner, the better the separation power. The Area Under the Curve (AUC) provides a single scalar measure of performance: AUC = 1 indicates perfect signal-background discrimination, while AUC = 0.5 corresponds to a random guessing.

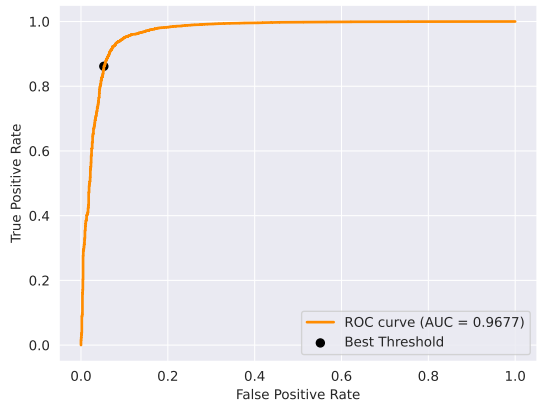


Figure 14: Receiver Operating Characteristic (ROC) curve for the neural network classifier.

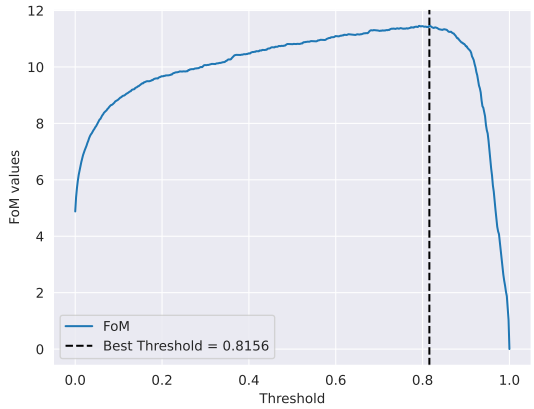


Figure 15: Scaled Figure of Merit as a function of NN output threshold. The maximum defines the optimal threshold for signal-background separation.

Using these standard metrics, the final accuracy (fraction of correctly classified events) is 94.11%, and the AUC is 96.77%.

6.5 Threshold Optimisation

To define the optimal threshold on the NN output, a standard Figure of Merit (FoM) is employed:

$$FOM = \frac{S}{\sqrt{S+B}}, \tag{12}$$

where S and B correspond to the signal and background event counts above a given threshold.

Since the number of MC signal events is limited only by computational resources, whereas the data sample size is fixed by the integrated luminosity, the FoM in (12) needs to be rescaled. The corresponding signal (f_S) and background (f_B) scale factors are defined as:

$$f_S = \frac{S_{\text{data}}}{S_{\text{MC}}}, \tag{13}$$

where S_{data} is the expected number of signal events in data and S_{MC} is that in the MC sample, and

$$f_B = \frac{B_{\text{SR}}}{B_{\text{SBL}} + B_{\text{SBR}}}, \tag{14}$$

where B_{SR} is the number of background events in the signal region, and B_{SBL} and B_{SBR} are those in the left and right sidebands, respectively. The necessary values are taken from the fits described in Sections 5.1 and 5.5.

Applying (13) and (14) to (12) yields the scaled FoM:

$$FOM_{\text{scaled}} = \frac{f_S \cdot S}{\sqrt{f_S \cdot S + f_B \cdot B}}. \tag{15}$$

Maximising this FoM as a function of the NN threshold (Fig.15) determines the optimal cut value for discriminating signal from background. The maximum is obtained at a threshold of 0.8156, which is subsequently used to classify events in the analysis.

7 Impact of ML-based Selection

In addition to the Machine Learning (ML) threshold cut, a best-candidate selection is applied to ensure only one B^0 candidate is retained per event. This is necessary because the reconstruction algorithm can produce multiple candidates in the same event due to different track combinations or combinatorial background. Allowing multiple candidates per event can artificially inflate the fitted yields and underestimate statistical uncertainties. To avoid this bias, only the candidate with the highest ML score is selected in each event, ensuring the most signal-like hypothesis is chosen.

The impact of this combined selection is evaluated by comparing the fitted signal and background yields, before and after the ML threshold and best-candidate choice. The post-selection yields are obtained using the same fitting procedures described in Sections 5.1 and 5.5, applied to both MC and data samples. The final binned likelihood fit to the data is shown in Fig. 16

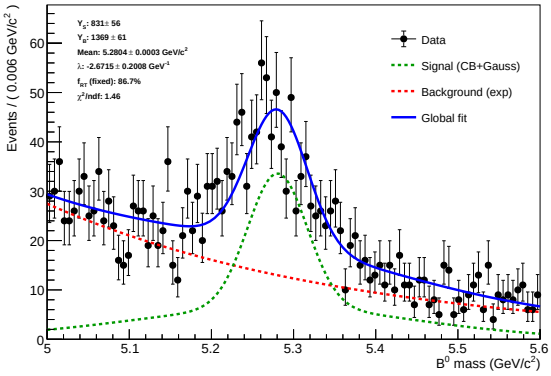


Figure 16: Binned likelihood fit to the B^0 mass distribution after ML threshold and best-candidate selection.

Table 2 summarises the fitted yields results, signal-to-background ratios and statistical significance ($S/\sqrt{S+B}$) for each approach. Here, the statistical significance, where S and B are the expected signal and background yields, quantifies the confidence level at which signal can be distinguished from background.

Table 2: Signal and background yields, signal-to-background ratios, and statistical significances, before (top) and after (bottom) applying ML threshold and best-candidate selection.

Y_S	Y_B	S/B	$S/\sqrt{S+B}$
595 ± 38	1871 ± 52	0.32	11.98
831 ± 56	1369 ± 61	0.61	17.73

These results demonstrate that the ML-based selection, together with the best-candidate criterion, significantly improve the purity of the sample, nearly doubling the S/B ratio and increasing the statistical significance by approximately 30%.

8 Conclusion

The decay channel $B^0 \to K^{*0}J/\Psi$ has been thoroughly analysed, using MC and CMS Run-3 data samples to discriminate signal from background. Initially, simple selection thresholds on variables such as vertex confidence level and flight length improved signal visibility in data, allowing the extraction of signal and background yields.

For the Multivariate Analysis (MVA), a Neural Network (NN) classifier was trained and optimised using a combination of correlation studies, SHAP importance - which resulted with 13 carefully selected features - and automated hyperparameter tuning with Optuna. The NN output, combined with a best-candidate selection, significantly enhances the purity of the selected sample. The resulting S/B ratio nearly doubles compared to the preselection case, and the statistical significance increases by approximately 30%, demonstrating the effectiveness of the ML-based approach.

The tools and methodology developed in this study can be applied in the future to the rare decay channel $B^0 \to K^{*0}\mu^+\mu^-$, providing a powerful framework for exploring flavour anomalies in dedicated datasets collected by CMS.

Acknowledgements

We would like to express our sincerest gratitude to our supervisors, Alessio Boletti and Nuno Leonardo, for their continuous guidance and encouragement throughout this project. Their availability and support were crucial for shaping the direction of this work and for deepening our understanding of both theoretical and experimental aspects of flavour physics. We are also deeply thankful to LIP for hosting the internship within a stimulating research environment, which provided us an opportunity to work closely with current topics in particle physics.

References

[1] LHCb Collaboration, Phys. Rev. Lett. **111**, 191801 (2013)

[2] LHCb Collaboration, Phys. Rev. Lett. **125**, 011802 (2020)

[3] CMS Collaboration, Phys. Lett. B **864**, 139406 (2025)

[4] Particle Data Group (2025), https://pdglive.lbl.gov/Viewer.action (accessed 2025-09-11)

[5] CMS, Journal of Instrumentation **3**, S08004 (2008)

[6] R.T. (CERN), *RooFit — A toolkit for modeling event distributions in ROOT*, CERN (2025), https://root.cern/manual/roofit/ (accessed: 2025-09-16)

[7] S.M. Lundberg, S.I. Lee, *A Unified Approach to Interpreting Model Predictions*, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Curran Associates, Inc., 2017), Vol. 30, available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[8] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, *Optuna: A Next-generation Hyperparameter Optimization Framework*, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, New York, NY, USA, 2019), KDD '19, pp. 2623–2631, available at: https://doi.org/10.1145/3292500.3330701