INVARIANT MASS PREDICTION USING NEURAL NETWORKS

Unlocking Particle Interactions with Deep Learning

25/06/2025

JON GIL ALVARO

LinkedIn: Jon Gil Alvaro

Contents

1.-Abstract

In this study, the intersection between high-energy particle physics and artificial intelligence is explored through the development of a neural network designed to predict the invariant mass of dielectron systems produced in electron-positron collision events.

A dataset provided by the CMS experiment at CERN is utilized, containing 100,000 events within the invariant mass range of 2–110 GeV. For each event, kinematic properties of both electrons are included, such as energy, momentum components, charge, pseudorapidity, and azimuthal angle.

After preprocessing steps were applied—including data cleaning, normalization, and feature selection—a fully connected feedforward neural network was trained to perform regression on the invariant mass (M) using 14 input features. The model was optimized using the Adam algorithm, and its performance was evaluated based on Mean Squared Error (MSE) and Mean Absolute Error (MAE).

It is shown that the model can capture the complex non-linear relationships present in collision data, achieving a reliable level of predictive accuracy. Through this work, the potential of machine learning techniques, such as Artificial Neural Networks, to support research efforts in modern physics is demonstrated.

2.-Introduction

The field of particle physics is devoted to understanding the fundamental constituents of matter and the interactions that govern them. At the forefront of this research is CERN, the European Organization for Nuclear Research, where large-scale experiments such as the Compact Muon Solenoid (CMS) are conducted to study high-energy collisions of subatomic particles.

A key quantity in analysing such collisions is the invariant mass, which represents a frame-independent measure of the combined energy and momentum of a particle system. In particular, the invariant mass of two electrons (or a dielectron system) is of central interest, as it can be used to identify resonances corresponding to known or hypothetical particles, such as the Z boson.

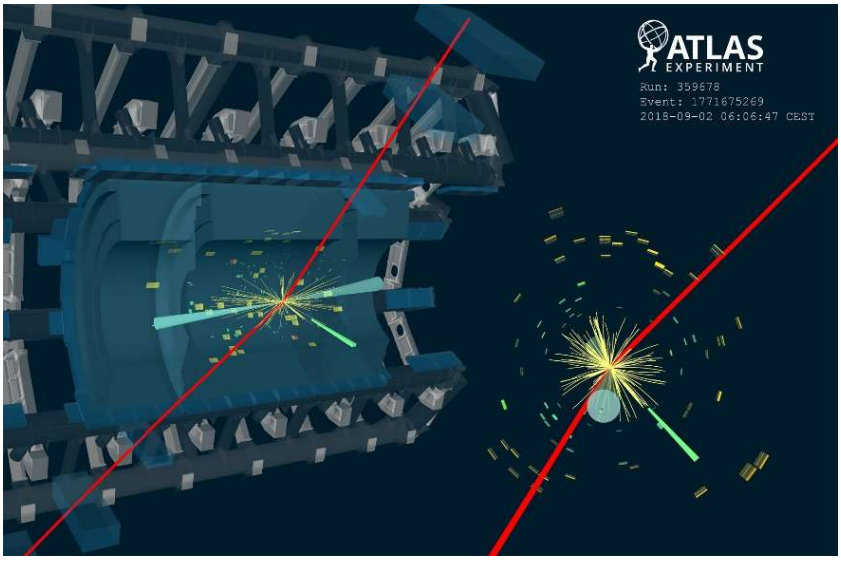


Image 1. Z boson

To facilitate education and outreach, a subset of CMS collision data has been made publicly available. This dataset consists of 100,000 electron-positron events and includes essential kinematic variables for each electron, such as energy, momentum components, charge, pseudorapidity, and azimuthal angle

In recent years, machine learning—particularly neural networks—has been increasingly employed in the physical sciences due to its ability to model complex, non-linear relationships in high-dimensional data. In this context, predictive models can be developed to approximate physical quantities based on observed parameters, offering new approaches to data analysis and interpretation.

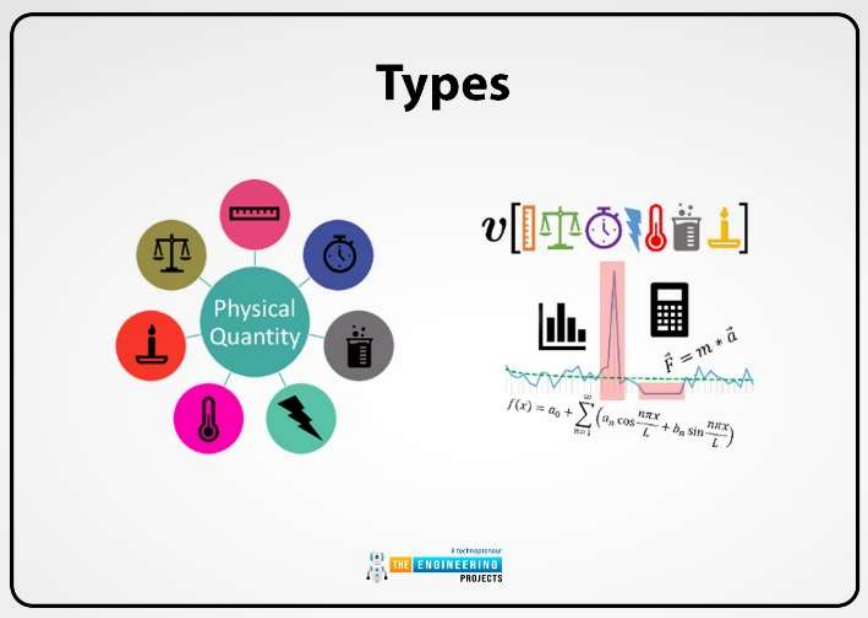


Image 2. ML and Physical Quantities

In the present work, a neural network is constructed and trained with the objective of predicting the invariant mass of dielectron events using only the kinematic properties of the individual electrons. By doing so, the potential of machine learning as a tool for both educational and research purposes in particle physics is further demonstrated.

3.-Dataset description

3.1.- Origin and Purpose of the Dataset

The dataset used in this study was provided by the CMS experiment at CERN and was specifically curated for educational and outreach purposes. It contains 100,000 dielectron collision events resulting from high-energy proton-proton collisions, with invariant mass values ranging from 2 to 110 GeV. Although simplified compared to datasets used in official CMS physics analyses, this dataset retains essential kinematic information that allows for meaningful exploration of physical properties and machine learning techniques.

3.2.- Variables Description

Each event in the dataset includes kinematic data for two electrons (labelled as electron 1 and electron 2). The following features are provided in the next Table 1:

Feature	Description
Run	Run number of the event
Event	Event number
E1, E2	Total energy of electrons 1 and 2 (GeV)
px1, px2	Momentum components along the x-axis for electrons 1 and 2 (GeV)
py1, py2	Momentum components along the y-axis for electrons 1 and 2 (GeV)
pz1, pz2	Momentum components along the z-axis for electrons 1 and 2 (GeV)
pt1, pt2	Transverse momentum of electrons 1 and 2 (GeV)
eta1, eta2	Pseudorapidity of electrons 1 and 2
phi1, phi2	Azimuthal angle (ϕ) of electrons 1 and 2 (radians)
Q1, Q2	Electric charge of electrons 1 and 2 (±1)
M	Invariant mass of the dielectron system (GeV) — **target variable**

Table 1. Feature description

3.2.1.- Explanation of Variables:

- *Energy (E1, E2):* The total energy of each electron in the event, measured in giga-electronvolts (GeV). It includes both rest mass energy and kinetic energy.
- *Momentum components (px, py, pz):* The three-dimensional components of the momentum vector of each electron, representing motion along the x, y, and z axes, respectively (in GeV/c). These are used to calculate other derived quantities, like transverse momentum and invariant mass.
- *Transverse momentum (pt1, pt2):* The component of the momentum perpendicular to the beam axis (usually the z-axis). It is defined as $\sqrt{{p_x}^2 + {p_y}^2 + {p_z}^2}$ and is important because it is less affected by the initial energy of the particles in the beam direction.
- *Pseudorapidity (eta1, eta2):* A spatial coordinate that describes the angle of a particle relative to the beam axis. It is preferred over the polar angle in high-energy physics due to its useful transformation properties under Lorentz boosts along the beam direction.
- *Azimuthal angle (phi1, phi2):* The angle of the particle's momentum in the transverse plane (x-y plane), measured in radians. It describes the particle's direction around the beam axis.
- *Charge (Q1, Q2):* The electric charge of the electron, which is either −1 (electron) or +1 (positron). Ensuring opposite charges (±1) is essential when selecting electron-positron (dielectron) events.
- *Invariant mass (M):* A scalar quantity calculated from the four-momenta of the two electrons. It is invariant under Lorentz transformations and is used to identify possible intermediate particles (like the Z boson) that may have decayed into the electron pair.

3.3.- Initial Data Processing

Before model development, several preprocessing steps were applied to ensure data quality and suitability for machine learning. First, all rows containing missing values in the target variable M were removed. Missing values were less than 0,01%, so the experiment would not be affected. Then, metadata columns such as Run and Event, as well as pt1 and pt2—which are derived from other momentum components—were excluded to avoid redundancy.

4.-Methodology

4.1.- Dataset Splitting

To ensure robust training and unbiased evaluation of the neural network, the dataset was divided into three distinct subsets:

- 85% for training, (84.297 samples)
- 14.8% for validation, (14.788 samples)
- 0.2% for final testing (200 samples)

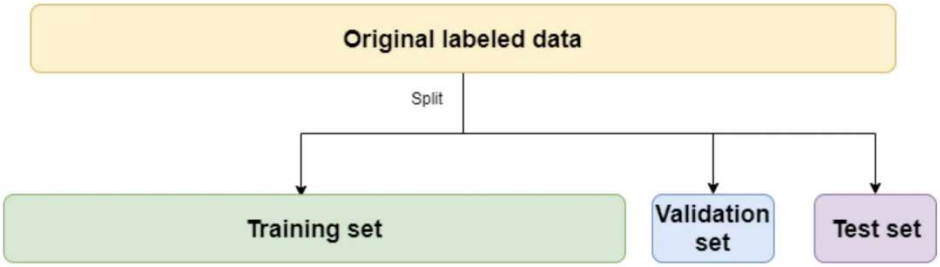


Image 3. Dataset Splitting

This particular split (see Image 3) was chosen to prioritize learning while still maintaining a meaningful validation set for monitoring the model's generalization during training. A very small fraction (0.2%) was held out as a final test set, which remained untouched during training and hyperparameter tuning. This separate test set was used exclusively for post-training evaluation, as new data cannot be simulated neither created.

4.2.- Feature Preprocessing

Before training, all input features were normalized using "StandardScaler", which transforms each feature to have zero mean and unit variance. This standardization was applied based only on the training data and then propagated to the validation and test sets to prevent data leakage.

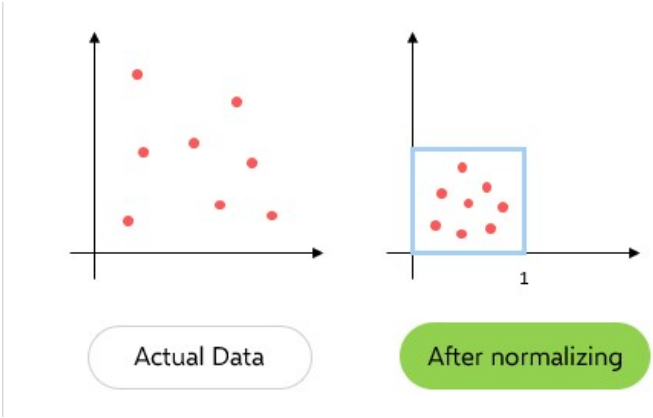


Image 4. Normalization of data

Feature scaling is particularly important in neural networks, as it ensures that all input features are on comparable numerical scales. Without normalization, features with larger numeric ranges (e.g., energy or momentum) could disproportionately influence the gradients, leading to inefficient or unstable learning.

4.3.- Model Architecture

A fully connected feedforward neural network was implemented, consisting of three hidden layers with 128, 64, and 32 neurons, respectively. All hidden layers employed the ReLU (Rectified Linear Unit) activation function, chosen for its simplicity and ability to mitigate the vanishing gradient problem in deep networks.

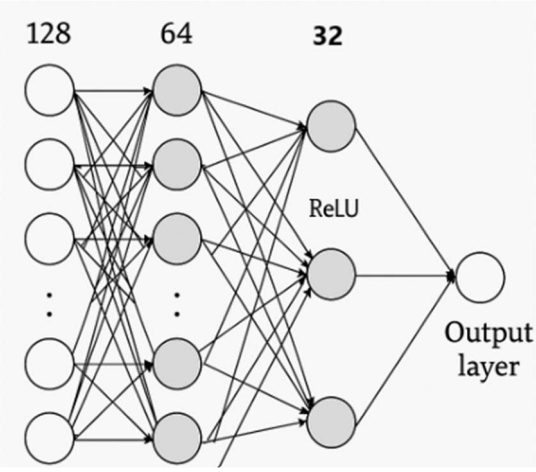


Figure 5. ANN architecture

To prevent overfitting, L2 regularization was applied to each layer. This technique penalizes large weights by adding a term proportional to the square of the weight magnitudes in the loss function. As a result, the model is encouraged to learn simpler, more generalizable patterns, rather than memorizing the training data.

The output layer consisted of a single neuron without activation, suitable for a regression task, where the goal is to predict a continuous variable: the invariant mass (M).

The loss function used was the Mean Squared Error (MSE), which penalizes larger errors more heavily and is standard for regression problems. In addition, the Mean Absolute Error (MAE) was tracked as a complementary metric, as it provides a more interpretable measure of average prediction error in physical units (GeV).

The model was trained using the Adam optimizer, a widely adopted adaptive optimization algorithm known for its fast convergence and stability in deep learning tasks.

4.4.- Training Procedure

The model was trained for 300 epochs with a batch size of 32. A validation set was used during training to monitor the model's performance at each epoch and to detect signs of overfitting or underfitting.

Training and validation loss values were recorded and plotted over time, providing insights into the convergence behaviour of the model. These learning curves served as diagnostic tools to assess whether the model had reached a satisfactory minimum or if further tuning was required.

5.-Results and Model Evaluation

5.1.- Quantitative Evaluation

To evaluate the model's predictive performance, the Mean Absolute Error (MAE) was computed on the held-out test set. This metric reflects the average absolute difference between predicted and true invariant mass values, expressed in GeV:

MAE on test set: 0.43 GeV

This value represents the typical deviation between the predicted invariant mass and its true value in the test events. Given the mass range (2–110 GeV), an error of this magnitude indicates that the model achieves reasonable precision within a few percent of the target range.

5.2.- Visualizations

To gain deeper insight into the model's behaviour, several diagnostic plots were generated. Each visualization was chosen to highlight a specific aspect of performance or model behaviour:

5.2.1.- Loss Curve (Training vs Validation)

A plot of training and validation Mean Squared Error (MSE) over epochs was used to assess convergence and generalization.

- Purpose: To observe whether the model is improving over time and whether overfitting or underfitting occurs.
- Interpretation: In the presented plot, the training loss stabilizes at an MSE value close to 1.0, indicating that the model has reached a consistent performance level during training without further significant improvement or degradation. Similarly, the validation loss remains close to 1.0, with only minor fluctuations throughout the epochs, which is expected behavior due to stochastic factors such as mini-batch updates and validation set variability.

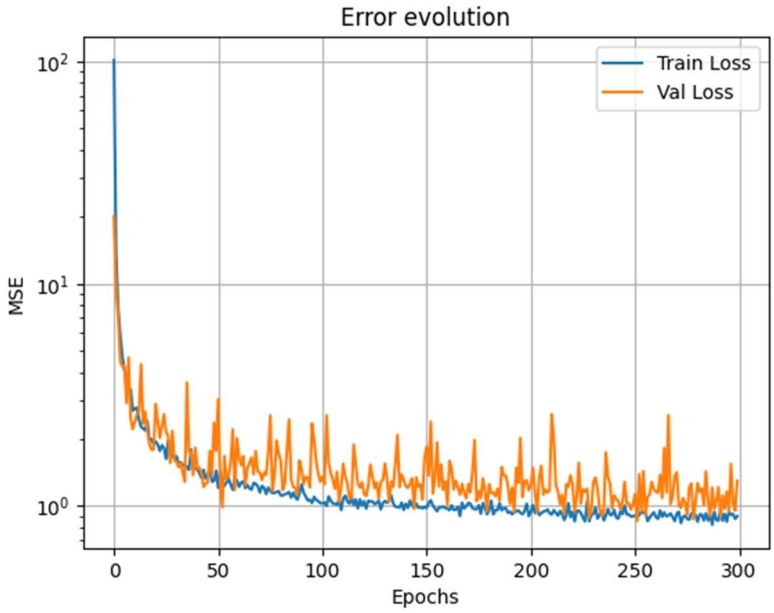
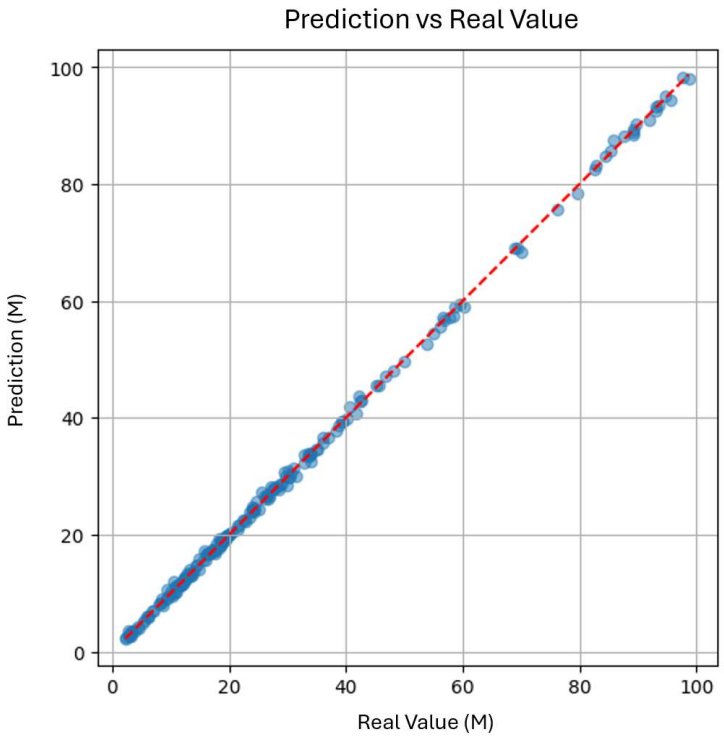


Figure 6. Loss Curve

5.2.2.- Prediction vs. Ground Truth

A scatter plot was generated comparing predicted invariant masses versus their true values, with a diagonal line indicating perfect prediction.

- Purpose: This visualization helps assess the accuracy of the model's predictions on an instance-by-instance basis. It provides an intuitive way to detect systematic errors, outliers, or regions where the model may struggle to generalize.
- Interpretation: In the resulting plot, most of the points are densely clustered along the diagonal line, indicating that the model's predictions are highly accurate and closely match the true invariant masses. The absence of significant systematic deviations from the diagonal suggests that the model is not exhibiting bias and is capturing the underlying data distribution effectively. The tight alignment also reflects good generalization and a strong correlation between predicted and actual values, reinforcing the conclusions drawn from the loss curves.



-

Figure 7. Prediction vs Ground truth

5.2.3.- Histogram of Absolute Error

The absolute differences between predictions and ground truth were plotted as a histogram.

- Purpose: This plot helps assess how large and how frequent the prediction errors are. It is useful for identifying the presence of outliers, the consistency of the model's performance, and potential regions of higher uncertainty.
- Interpretation: The histogram shows that most absolute errors are concentrated in the range from 0 to approximately 0.75, indicating that the model generally makes small and consistent prediction errors. This concentration suggests a high level of predictive accuracy for many cases. The distribution then tapers off, with significantly fewer errors

observed in the range from 1.0 to 1.75, which may correspond to occasional harder-to-predict events or noise in the data.

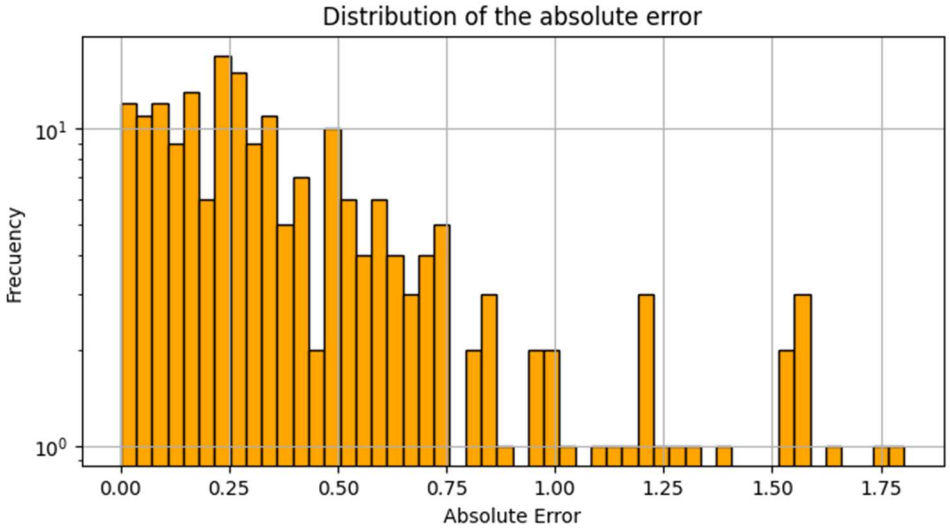


Figure 8. Histogram of Absolute Error

5.2.4.- Histogram of Relative Error

A histogram of the relative errors—computed as the absolute error divided by the true value—was plotted to evaluate the model's performance in a scale-invariant manner.

Purpose: Relative error provides a normalized view of prediction accuracy across a wide dynamic range of target values (e.g., invariant masses spanning from 2 to 110 GeV). This allows for fairer evaluation of performance, since a fixed absolute error has a different impact depending on the scale of the target value.

Interpretation: The histogram reveals that the vast majority of relative errors fall within a very narrow range between 0% and 0.1%, demonstrating that the model maintains a high level of precision across the entire mass spectrum. This tightly concentrated distribution indicates that even in cases where absolute errors appear larger (as seen in the previous section), they correspond to higher-mass instances and are proportionally small when normalized.

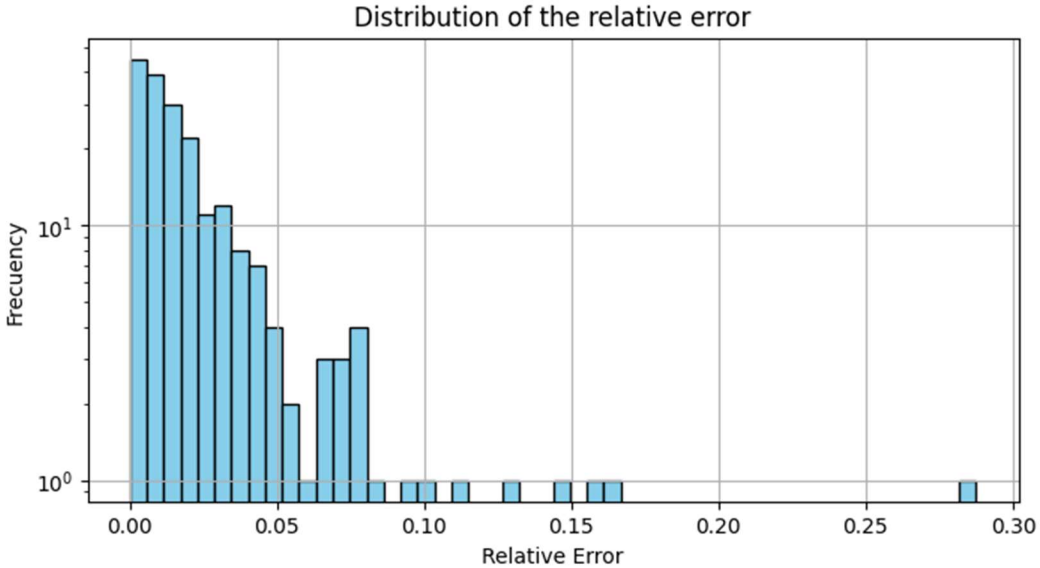


Figure 9. Histogram of Relative Error

5.2.5.- Absolute Error vs. True Mass

A scatter plot was created showing absolute error as a function of the true invariant mass.

- Purpose: This plot helps identify specific regions (e.g., low or high mass ranges) where the model may struggle or excel. It is particularly useful for detecting patterns such as systematic increases in error, which may arise from data imbalance, reduced input sensitivity, or intrinsic complexity in certain regions.
- Interpretation: The majority of data points are concentrated within a bounded region—from 0 to 40 GeV on the x-axis (true mass) and from 0 to approximately 0.75 on the y-axis (absolute error). This indicates that for a large portion of the dataset, particularly in the low-to-mid mass range, the model maintains low and consistent error levels. There are no strong visible trends of increasing error with increasing mass within this core region, suggesting stable performance. Larger absolute errors seen in prior plots appear to be associated with higher-mass events but as shown earlier, are relatively small when considered in relative terms. Overall, this plot confirms that the model performs robustly across most of the spectrum, with no major regions of systematic underperformance.

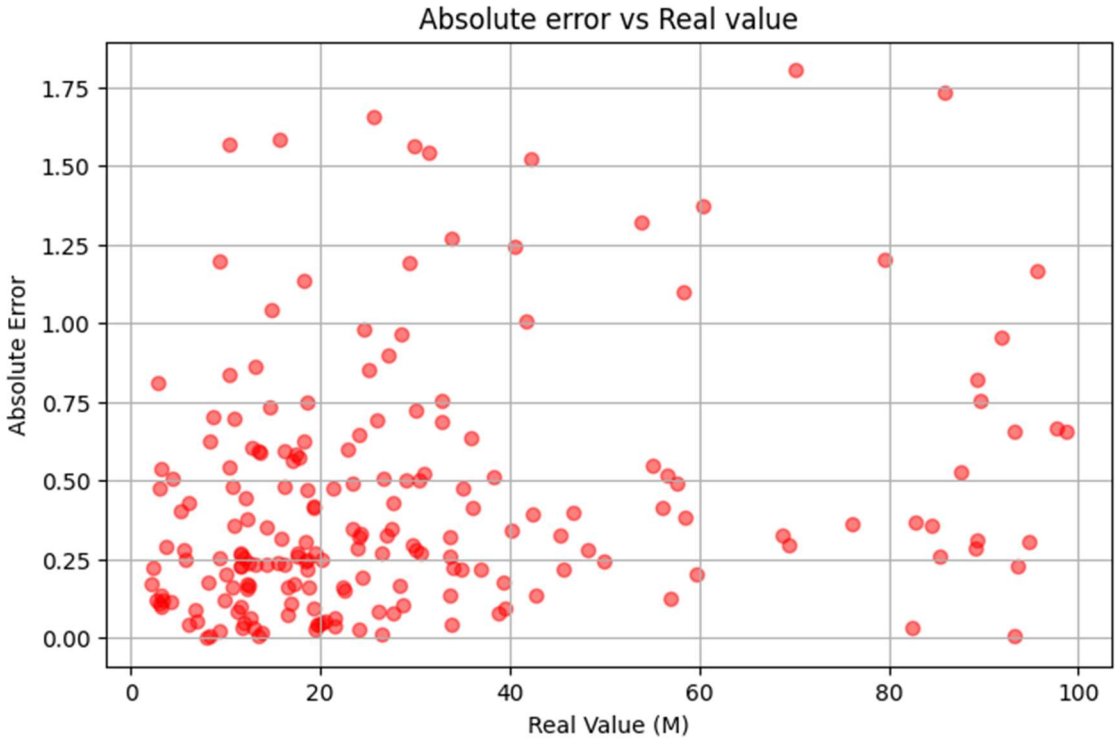


Figure 10. Absolute error scatter plot

6.-Correlation Analysis Between Variables

To gain further insight into the relationships among the input features and the target variable, a Pearson correlation matrix was computed using the entire dataset. This matrix quantifies the strength and direction of linear relationships between each pair of variables.

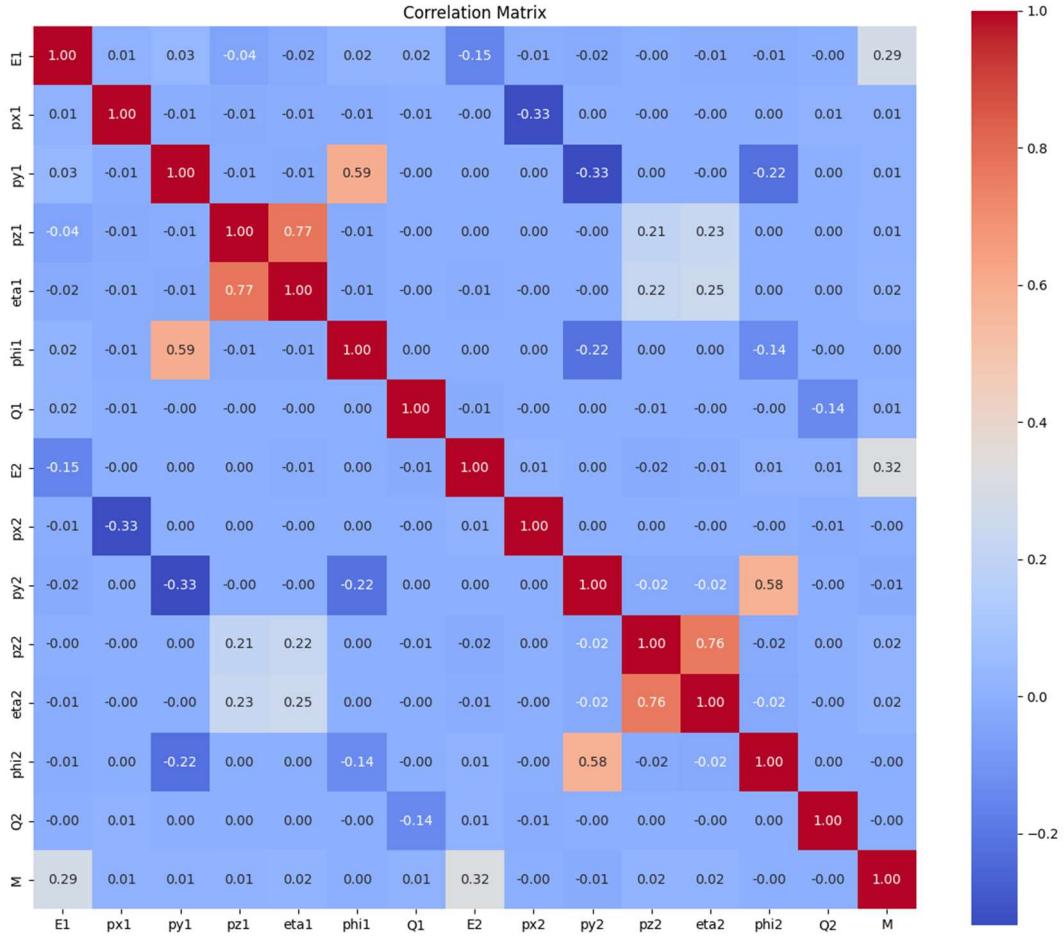


Figure 11. Correlation Matrix

6.1.- Overall Correlation Structure

The correlation matrix revealed several strong interdependencies among kinematic variables (e.g., momentum components px, py, pz), particularly within the same electron (e.g., phi1 and etha1).

Several features exhibited high internal correlations, indicating potential redundancy or multicollinearity, which could affect model interpretability but are generally well-handled by neural networks.

6.2.- Relationship to the Target Variable (Invariant Mass)

Surprisingly, the invariant mass M showed only moderate linear correlations with the total energies of the electrons:

- E1: 0.30
- E2: 0.36

All other features exhibited very weak or near-zero correlations with M, suggesting that the relationship between input variables and the invariant mass is non-linear in nature.

This finding aligns with physical expectations: the invariant mass is calculated from the four-momenta of the electron pair via a specific relativistic formula, which is non-linear and cross-dependent on all momentum components. As such, linear correlation is not sufficient to capture the predictive potential of individual features in this context.

6.3.- Implications and Interpretation

The moderate correlations of E1 and E2 with M indicate that total energy contributes meaningfully to the mass, as expected from relativistic kinematics. However, the modest correlation values confirm that no single variable alone is strongly predictive of the invariant mass.

The neural network's ability to accurately predict M despite low individual correlations highlights its capacity to capture complex, multi-variable interactions that govern the underlying physics.

Features with strong mutual correlations (e.g., between momentum components) could be seen as redundant, but in practice they may still provide valuable signal when used together in a non-linear model.

7.-Conclusions

In this study, a supervised machine learning approach was applied to the problem of predicting the invariant mass of electron pairs using kinematic variables derived from high-energy collision data provided by CERN. A fully connected feedforward neural network was trained on an educational CMS dataset containing 100,000 dielectron events within a mass range of 2–110 GeV.

Several key findings were obtained:

- The model achieved a mean absolute error of 0.43 Gev on the test set, demonstrating that the invariant mass can be predicted with reasonable precision using only the measured features of the individual electrons.
- The close alignment of training and validation losses suggests that the model generalizes well to unseen data and is not overfitting. The absence of a widening gap between the two curves, combined with their stability, implies that the learning process has converged effectively, and the model is neither underfitting nor overfitting.
- Relative error confirms that the model's performance does not degrade for higher mass values and that its predictive accuracy scales appropriately with the magnitude of the target. The use of relative error thus reinforces the conclusion that the model is not only accurate in absolute terms but also robust across varying scales.
- Despite the weak linear correlation between the invariant mass and any single input variable (with the highest correlations being 0.30 for E1 and 0.36 for E2), the neural network was able to learn the underlying non-linear relationships among the input variables, which are governed by the principles of relativistic kinematics.

The results obtained here support the viability of artificial intelligence techniques—specifically deep learning—as powerful tools for extracting physical insights and making accurate predictions from particle physics data. While traditional analysis relies on explicit formulae and assumptions, machine learning models can learn complex relationships directly from data, even when those relationships are non-trivial or not easily expressed analytically.

This approach offers promising applications not only in educational contexts but also in real-world experimental analyses, where large volumes of multidimensional data are generated and where nuanced, high-precision predictions are required.