VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

An Autonomous Institute Affiliated to University of Mumbai

Department of Computer Engineering

Project Report on

Electron energy analysis of high granularity calorimeter using ML methods

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2023-24

Submitted by

Aayush Shribatho (D17A , Roll no - 64)

Sahil Salunkhe (D17A , Roll no - 58)

Abhayvir Singh (D17A , Roll no - 01)

Hitesh Ramrakhyani (D17A , Roll no - 55)

Project Mentor

Dr. Sharmila Sengupta

(2023-24)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

An Autonomous Institute Affiliated to University of Mumbai

Department of Computer Engineering

Certificate

This is to certify that ***Aayush Shribatho (D17A 64), Abhayvir Singh(D17A 01), Sahil Salunkhe(D17A 58) and Hitesh Ramrakhyani(D17A 55)*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on " ***Electron energy analysis of high granularity calorimeter using ML methods***" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Dr. Sharmila Sengupta*** in the year 2023-24 .

Programme Outcomes	Grade
PO1, PO2, PO3, PO4, PO5, PO6, PO7, PO8, PO9, PO10, PO11, PO12, PSO1, PSO2	

Date:

Project Guide:

टाटा मूलभूत अनुसंधान संस्थान
TATA INSTITUTE OF FUNDAMENTAL RESEARCH
होमी भाभा रोड, कोलाबा, मुंबई – ४०० ००५.
Homi Bhabha Road, Colaba, Mumbai - 400 005.
परमाणु ऊर्जा विभाग की स्वायत्त संस्था
भारत सरकार एवं समविश्वविद्यालय
An Autonomonus Institution of the Department of Atomic Energy
Government of India and Deemed University

दूरभाष / Telephone : +91 22 2278 2000
फैक्स / Fax : +91 22 2280 4610 / 11

वेबसाईट / Website : www.tifr.res.in

Date: 05/8/2023

Letter of Permission

This is to certify that following Final year students of Department of Computer Engineering of Vivekanand Education Society's Institute of Technology, Chembur, are working on a TIFR project titled "**CMS physics with background noise using GNN**", under the guidance of **Dr.Sharmila Sengupta and Mrs. Sunita Sahu** for the academic year 2023-24.

1. **Aayush Shribatho**
2. **Sahil Salunkhe**
3. **Hitesh Ramrakhyani**
4. **Abhayvir Singh**

We will provide all technical assistance to students required during the completion of the project. The progress seminars and meetings will be regularly conducted to take feedback.

Dr. Shashikant Dugad,

Professor, Department of High energy Physics,

Tata Institute of Fundamental Research, Mumbai

Project Report Approval
For
B. E (Computer Engineering)

This project report entitled ***Electron energy analysis of high granularity calorimeter using ML methods*** by ***Aayush Shribatho, Abhayvir Singh, Sahil Salunkhe and Hitesh Ramrakhyani*** is approved for the degree of ***B.E. Computer Engineering***.

Internal Examiner

External Examiner

Head of the Department

Principal

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---------- ----------

(Aayush Shribatho - D17A/64) (Abhayvir Singh - D17A/01)

---------- ----------

(Sahil Salunkhe - D17A/58) (Hitesh Ramrakhyani - D17A/55)

Date:

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Dr. Sharmila Sengupta, Associate Professor (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr. (Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair,** for giving us this valuable opportunity to do this project.

We also express our gratitude towards **Dr. Shashi Dugad** for offering and guiding us throughout the project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement at several times.

Computer Engineering Department

COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solutions for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.

CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop a professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

Index

Abstract

Physicists and engineers at CERN (Geneva) are studying the basic constituents of matter. The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator. Inside the accelerator, two high-energy particle beams travel at close to the speed of light before they are made to collide. Detectors observe and record the results of these collisions.Large Hadron Collider (LHC) lies in a tunnel of 27 kilometers (17 mi) in circumference and as deep as 175 meters (574 ft) beneath the France–Switzerland border near Geneva.. LHC is a machine which collides protons with an energy of 13 TeV. About 30 protons smash each other 40 million times a second .About a billion collision per second take place at the LHC and About 1 Higgs Boson is observed every second

The CMS experiment at CERN has a broad physics programs. Various particles and their natures are detected inside the CMS.The CMS is a multi-layered, general purpose detector designed to capture and measure various types of particles produced in high-energy collisions within the LHC. It is designed to observe any new physics phenomena that the LHC might reveal.HGCAL is a specialized type of particle detector used in high-energy physics experiments. Its primary purpose is to measure the energy of particles and to enhance particle identification. HGCAL is characterized by extremely fine segmentation into small cells or pixels HGCAL is optimized for measuring electromagnetic showers produced by electrons and photons.

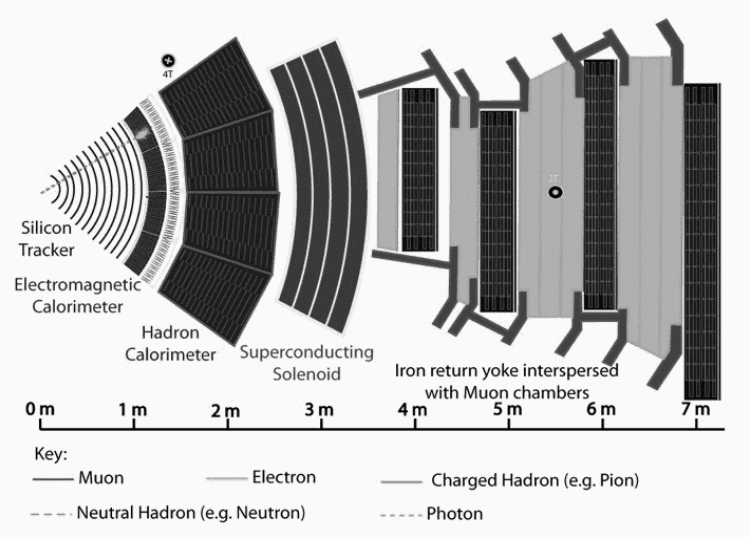
The project aims to leverage the power of GCN, a deep learning architecture designed to process graph-structured data, to perform regression analysis and predict the energy levels of electrons.

1.Introduction

The introduction section provides brief information about the project. It also includes the motivation behind the project and the drawbacks behind the existing system. A detailed problem definition description is provided which discusses the problem statement in detail. Followed by there is a subsection on relevance of the project. Finally, last but not least, is the subsection on Methodology used.

1.1 Introduction

The CMS experiment at CERN has a broad physics program ranging from studying the Standard Model to searching for extra dimensions and particles that could make up dark matter. In the CMS experiment at CERN, Geneva, a large number of HGCAL sensor modules were fabricated in advanced laboratories around the world. Various particles and their natures are detected inside the CMS large hadron collider. The Large Hadron Collider (LHC) is the world's most powerful and largest particle accelerator. The LHC consists of a 27-kilometer ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way. It accelerates protons to nearly the velocity of light and then collides them at four locations around its ring. To face the new challenges of the LHC, Physicists at the CMS collaboration are working on a completely new calorimeter, the High-Granularity Calorimeter which will deliver 10 times more collisions to its experiments to be installed in the endcaps of the detector. Each particle that emerges from an LHC collision is like a piece of a puzzle, with some of these pieces breaking up further as they travel away from the collision. Each leaves a trace in the detector and CMS’s job is to gather up information about every one - perhaps 20, 100 or even 1000 puzzle tracks - so that physicists can put the jigsaw back together and see the full picture of what happened at the heart of the collision. The inner tracking system of CMS is designed to provide a precise and efficient measurement of the trajectories of charged particles emerging from the LHC collisions, as well as a precise reconstruction of secondary vertices. It surrounds the interaction point and has a length of 5.8m and a diameter of 2.5m. Deep learning has recently been effectively applied to image recognition. Identification of particles and their trails can be done from photos using a combination of deep learning and image processing technologies.



1.2 Motivation for the project

The project domain is machine learning and the main focus area is regression and graph convolutional networks. Our model predicts the energy of electrons using a GCN model. This task was previously done by using image processing which was a tedious task as generating quality information from images at the LHC is difficult which also affects the accuracy while predicting the event energy. Our GCN model does the same work efficiently and with more accuracy using neural networks and deep learning.

1.3 Problem Definition

This project is aimed at detection of physical objects in the CMS experiment at CERN, Geneva, CMS is performing at its peak, about one billion proton-proton interactions will take place every second inside the detector. There is no way that data from all these events could be read out, and even if they could, most would be less likely to reveal new phenomena. It is very important to identify the objects inside the collider. We therefore need a “trigger” that can select the potentially interesting events, such as those which will produce the Higgs particle, and reduce the rate to just a few hundred “events” per second, which can be read out and stored on computer disks for subsequent analysis.

Various particles like electrons, muons etc, collide inside the large hadron collider and we need to calculate their energy for gaining meaningful insights from them. When high-energy protons collide in the CMS detector it initiates electron showers. These showers consist of a cascade of secondary particles, including photons (gamma rays), electrons, and positrons. Out of these particles, we will specifically focus on the electrons and calculate their energy using regression and deep learning methods.

1.4 Existing system

- Manually detecting the particles with just images is very difficult and tedious as the output number of images is in lakhs.
- The lighting of the output images is also varying which makes it difficult to generate any consistent output.
- Automation processes have lesser accuracy and are aimed at trying to detect a particle based on size.
- Software's like YOLO can detect images on the basis of a limited number of colors.
- An image having a mixture of colors cannot be detected and classified easily.

.

1.5 Lacuna of the existing systems

- High Volume of Images: The process involves dealing with a large volume of images, possibly in the hundreds of thousands or more.
- Manual Inspection: Each image must be manually inspected by a human operator to detect particles, which is a time-consuming and labor-intensive task.
- Error Prone: Human error is inevitable, and even the most careful inspection can result in missed particles or false positives.
- Resource Intensive: Manually detecting particles consumes human resources that could be allocated to other tasks, potentially slowing down overall workflow and productivity.

1.6 Relevance of the Project

The project is aimed at developing a deep learning model that will be able to able to calculate the energy of electrons generated in the large hadron collider and will be used for research at the CMS experiment facility at CERN. This project, if successful, will be deployed by TIFR in their research on the CMS experiment.

Manually detecting the particles with just images is very difficult and tedious as the output number of images is in lakhs. The lighting of the output images is also varying which makes it difficult to generate any consistent output. So we propose a DL model which will help us to calculate the energy of electrons when a dataset is provided to the model. The dataset is generated at the CERN using the CMSSW software and it is stored in root software.

2. Literature Survey

Paper 1: HGCAL: a High-Granularity Calorimeter for the endcaps of CMS at HL-LHC
https://sci-hub.se/10.1088/1748-0221/12/01/c01042

Abstract:

Calorimetry at the High Luminosity LHC (HL-LHC) faces two enormous challenges, particularly in the forward direction: radiation tolerance and unprecedented in-time event pileup. To meet these challenges, the CMS experiment has decided to construct a High Granularity Calorimeter (HGCAL), featuring a previously unrealized transverse and longitudinal segmentation, for both electromagnetic and hadronic compartments. This will facilitate particle-flow-type calorimetry, where the fine structure of showers can be measured and used to enhance particle identification, energy resolution and pileup rejection. The majority of the HGCAL will be based on robust and cost-effective hexagonal silicon sensors with ≃ 1 cm2 or 0.5 cm2 hexagonal cell size, with the final five interaction lengths of the hadronic compartment being based on highly segmented plastic scintillator with on-scintillator SiPM readout. We present an overview of the HGCAL project, including the motivation, engineering design, readout/trigger concept and simulated performance.

Inference:

The implementation of the High Granularity Calorimeter (HGCAL) by the CMS experiment at the High Luminosity Large Hadron Collider (HL-LHC) is aimed at addressing significant challenges posed by radiation tolerance and unprecedented in-time event pileup, particularly in the forward direction. By incorporating previously unrealized levels of transverse and longitudinal segmentation for both electromagnetic and hadronic compartments, the HGCAL enables particle-flow-type calorimetry. This innovative approach allows for the measurement of fine shower structures, leading to improved particle identification, energy resolution, and pileup rejection capabilities. Utilizing robust and cost-effective hexagonal silicon sensors for the majority of the detector, supplemented by highly segmented plastic scintillators for the final five interaction lengths of the hadronic compartment, the HGCAL project aims to enhance the

overall performance of calorimetry at the HL-LHC, thus facilitating a deeper understanding of particle interactions and phenomena.

Paper 2: THE CMS HGCAL DETECTOR FOR HL-LHC UPGRADE

https://arxiv.org/pdf/1708.08234.pdf

Abstract:

The High Luminosity LHC (HL-LHC) will integrate 10 times more luminosity than the LHC, posing significant challenges for radiation tolerance and event pileup on detectors, especially for forward calorimetry, and hallmarks the issue for future colliders. As part of its HL-LHC upgrade program, the CMS collaboration is designing a High Granularity Calorimeter to replace the existing endcap calorimeters. It features unprecedented transverse and longitudinal segmentation for both electromagnetic (ECAL) and hadronic (HCAL) compartments. This will facilitate particle-flow calorimetry, where the fine structure of showers can be measured and used to enhance pileup rejection and particle identification, whilst still achieving good energy resolution. The ECAL and a large fraction of HCAL will be based on hexagonal silicon sensors of 0.5 - 1 cm2 cell size, with the remainder of the HCAL based on highly-segmented scintillators with SiPM readout. The intrinsic high-precision timing capabilities of the silicon sensors will add an extra dimension to event reconstruction, especially in terms of pileup rejection. An overview of the HGCAL project is presented, covering motivation, engineering design, readout and trigger concepts, and performance (simulated and from beam tests).

Inference:

The upgrade program for the High Luminosity Large Hadron Collider (HL-LHC), specifically addressing challenges related to radiation tolerance and event pileup on detectors, underscores the significance of advancing detector technologies for future colliders. The design of the High Granularity Calorimeter by the CMS collaboration represents a proactive response to these challenges, aiming to replace existing endcap calorimeters with a system featuring unprecedented levels of transverse and longitudinal segmentation. By enabling particle-flow calorimetry and leveraging the fine structure of showers, the new calorimeter promises enhanced pileup rejection and particle identification capabilities while maintaining good energy resolution.

The utilization of hexagonal silicon sensors for the majority of the detector, complemented by highly-segmented scintillators, showcases a multi-faceted approach to achieving high-performance detection. Furthermore, the integration of high-precision timing capabilities provided by silicon sensors presents an additional dimension to event reconstruction, particularly beneficial for pileup rejection. Overall, the HGCAL project represents a comprehensive effort to address the evolving demands of particle physics experiments at the HL-LHC and beyond.

Paper 3: Electromagnetic showers beyond shower shapes
https://sci-hub.se/https://www.sciencedirect.com/science/article/abs/pii/S0168900219312999

Abstract:
Correctly identifying the nature and properties of outgoing particles from high energy collisions at the Large Hadron Collider is a crucial task for all aspects of data analysis. Classical calorimeter-based classification techniques rely on shower shapes — observables that summarize the structure of the particle cascade that forms as the original particle propagates through the layers of material. This work compares shower shapebased methods with computer vision techniques that take advantage of lower level detector information. In a simplified calorimeter geometry, our DenseNet-based architecture matches or outperforms other methods on e +-γ and e +-π + classification tasks. In addition, we demonstrate that key kinematic properties can be inferred directly from the shower representation in image format.

Inference:
The study highlights the significance of accurately identifying outgoing particles from high-energy collisions at the Large Hadron Collider (LHC) for comprehensive data analysis. Traditionally, classification techniques rely on analyzing shower shapes, which summarize the cascade structure formed by particles traversing detector materials. However, this research introduces a novel approach by comparing shower shape-based methods with computer vision techniques that leverage detailed detector information at a lower level. In a simplified calorimeter geometry, the DenseNet-based architecture showcased in the study either matches or surpasses other methods in classifying electron-positron and electron-positron-pion events. Furthermore, the research demonstrates the feasibility of inferring key kinematic properties

directly from the shower representation in image format, opening avenues for more efficient particle identification and analysis in high-energy collision experiments.

Paper 4: The large hadron collider

https://www.sciencedirect.com/science/article/abs/pii/S0146641012000695

Abstract:

The Large Hadron Collider (LHC) is the world's largest and most energetic particle collider. It took many years to plan and build this large complex machine which promises exciting, new physics results for many years to come. We describe and review the machine design and parameters, with emphasis on subjects like luminosity and beam conditions which are relevant for the large community of physicists involved in the experiments at the LHC. First collisions in the LHC were achieved at the end of 2009 and followed by a period of a rapid performance increase. We discuss what has been learned so far and what can be expected for the future.

Inference:

The Large Hadron Collider (LHC) stands as a monumental achievement in scientific engineering, representing the culmination of years of meticulous planning and construction. As the largest and most energetic particle collider in the world, its significance transcends mere scientific curiosity, holding the promise of unlocking groundbreaking discoveries in physics for years to come. This review delves into the intricate design and parameters of the LHC, shedding light on crucial aspects such as luminosity and beam conditions that are pivotal for the vast community of physicists engaged in experiments at the facility. Since its inception, the LHC has undergone a remarkable journey, marked by the achievement of first collisions in 2009 followed by a period of rapid performance enhancement. Through this exploration, we reflect on the invaluable lessons learned thus far and anticipate the exciting prospects that lie ahead on the horizon of particle physics research.

Paper 5: Response of a CMS HGCAL silicon-pad electromagnetic calorimeter prototype to 20–300 GeV positrons

https://iopscience.iop.org/article/10.1088/1748-0221/17/05/P05022/meta

Abstract:

The Compact Muon Solenoid (CMS) collaboration is gearing up for the High Luminosity-LHC (HLLHC) by upgrading its endcap calorimeters with a new high-granularity calorimeter (HGCAL). This HGCAL is designed to tackle the increased radiation levels and overlapping events (pileup) expected during HLLHC operations compared to current LHC conditions. The novel design incorporates hexagonal multipad large-area silicon sensors and plastic scintillator tiles, offering enhanced reconstruction of physics objects with improved tolerance to radiation damage. The high granularity enables precise particle-flow measurements extending from the tracker into the calorimeter, facilitating effective subtraction of energy from pileup events and improving energy resolution, especially for merged jets and narrow jets from specific production modes like Higgs boson decay. Prototype development began in 2015, with beam tests conducted in 2016 and 2018 to validate the design and assess performance.

Inference:

The CMS collaboration is preparing for the HLLHC era by upgrading the endcap calorimeters with the HGCAL, designed to withstand higher radiation levels and cope with increased pileup events. The novel design features hexagonal silicon sensors and plastic scintillator tiles, promising better physics object reconstruction and radiation damage tolerance. The high granularity of the detector enables precise energy measurements and efficient pileup event subtraction, enhancing performance in detecting various particle jets. Prototype testing has shown promising results, setting the stage for further development and integration into the CMS detector system.

Paper 6: The Graph Neural Network Model

https://ieeexplore.ieee.org/abstract/document/4700287

Abstract:

The paper discusses the representation and processing of structured data, particularly graphs, in various application areas such as proteomics, image analysis, software engineering, and natural language processing. It distinguishes between two types of machine learning applications: graph-focused, where the function maps a graph to a vector independently of nodes, and node-focused, where the function depends on individual node properties. Traditional approaches preprocess graph data into simpler representations, potentially losing valuable information. However, recent techniques aim to preserve graph structure during processing. The paper introduces a novel neural network model called a graph neural network (GNN), which unifies existing models for processing both graph and node-focused applications. GNNs utilize an information diffusion mechanism to process graphs and can handle various types of graphs without preprocessing.

Inference:

The paper addresses the challenges of representing and processing structured data, particularly graphs, in machine learning applications. It highlights the limitations of traditional preprocessing approaches and introduces a novel neural network model, the graph neural network (GNN), which can handle both graph and node-focused applications while preserving the graph structure. By utilizing an information diffusion mechanism, GNNs offer a unified framework for processing diverse types of graphs without the need for preprocessing. This approach has the potential to enhance the performance of machine learning algorithms on graph-structured data in various domains.

Paper 7: The CMS HGCAL trigger data receiver

https://iopscience.iop.org/article/10.1088/1748-0221/19/01/C01049/meta

Abstract:

The article presents the development and testing of a prototype receiver for the High Granularity Calorimeter (HGCAL) endcap front end, as part of the CMS Phase-2 upgrade. Implemented using the Serenity ATCA platform, the receiver firmware was designed to unpack data from the front-end endcap trigger concentrator ASIC (ECON-T) and evaluate its performance and stability. Integrated with prototype data acquisition (DAQ) firmware and ancillary blocks, the system aims to generate triggers using ECON-T data, process scintillator triggers, and evaluate the complete HGCAL vertical slice. Data is read out using custom 10G UDP links and upgraded to CMS DTH system at 25 Gbps. The system achieved prototype Trigger Primitive Generator (TPG) Stage-1 and DAQ path readout, delivering around 20 TB of data containing physics events at an average trigger rate of about 100 kHz.

Inference:

The development and testing of the CMS HGCAL trigger data receiver represent significant progress towards the CMS Phase-2 upgrade. By successfully implementing a prototype receiver using advanced hardware and firmware, the study demonstrates the feasibility of unpacking data from the front-end endcap trigger concentrator ASIC and evaluating the performance of the entire HGCAL vertical slice. The integration with prototype DAQ firmware and ancillary blocks allows for comprehensive testing and validation of trigger generation, scintillator trigger processing, and data readout. The achieved data readout rates and trigger rates indicate promising performance capabilities of the system, paving the way for further advancements in the CMS experiment's capabilities for particle physics research.

Paper 8: ROOT: A Data Analysis and Data Mining Tool from CERN
https://www.casact.org/sites/default/files/database/forum_08wforum_kumar_tripathi.pdf

Abstract:

This paper introduces ROOT, an open-source data analysis framework developed at CERN, and explores its suitability for data analysis tasks in various fields, including high-energy physics and insurance analytics. ROOT is designed to efficiently handle large-scale data analysis by storing data in a hierarchical, object-oriented database, which is highly compressed and machine-independent. Unlike relational databases, which are optimized for transactional systems, ROOT excels at segmenting and analyzing data across multiple dimensions. This paper highlights ROOT's capabilities in multi-dimensional histograms, curve fitting, modeling, and simulation, making it a valuable tool for predictive modeling and ad hoc data analysis tasks. ROOT's optimal data storage and retrieval methods ensure scalability and efficient memory usage, enabling high-performance computing even on standard PCs.

Inference:

ROOT, developed at CERN, is a powerful and efficient data analysis framework suitable for various fields, including high-energy physics and insurance analytics.

Unlike relational databases, ROOT's hierarchical, object-oriented database allows for efficient storage and retrieval of large-scale data, making it ideal for segmenting and analyzing data across multiple dimensions.

ROOT's built-in tools for multi-dimensional histograms, curve fitting, modeling, and simulation enhance its usability for predictive modeling and ad hoc data analysis tasks.

ROOT's optimal data storage and retrieval methods ensure scalability and efficient memory usage, enabling high-performance computing even on standard PCs.

ROOT's interactive computing capabilities enable efficient access to subsets of data without impacting the rest of the dataset, making it suitable for interactive data exploration and analysis.

Paper 9: “Deep Learning in Particle Physics" by Kim Albertsson et al**.**
(https://arxiv.org/abs/1806.11484)

Abstract:

Particle physics experiments, notably those at the Large Hadron Collider (LHC), produce vast and intricate datasets that defy traditional analysis methods. Deep learning has emerged as a powerful tool in tackling these challenges, surpassing conventional techniques by effectively handling high-dimensional and complex data. This paper provides a concise overview of deep learning's application in LHC data analysis, catering to readers familiar with high-energy physics but less so with machine learning. It discusses the significance of machine learning in addressing the complexities of LHC data, outlines basic neural network concepts, reviews key applications of deep learning in LHC data analysis, and explores future directions and concerns. Overall, the paper underscores deep learning's transformative impact on particle physics data analysis, offering insights into its potential for advancing our understanding of fundamental physics phenomena.

Inference:

Deep learning has revolutionized the field of particle physics, particularly in the analysis of data from experiments such as those conducted at the Large Hadron Collider (LHC). Traditionally, high-energy physics (HEP) data analysis relied on sequential decision-making and statistical analysis based on single observed quantities, limiting its effectiveness in handling high-dimensional data. However, the advent of deep learning has provided a paradigm shift, enabling the exploitation of complex, multi-variable data in ways previously unattainable. By leveraging techniques like artificial neural networks, deep learning algorithms can effectively navigate the intricate landscape of LHC data, surpassing the capabilities of traditional methods, especially as data dimensionality grows. This review underscores the crucial role of machine learning, particularly deep learning, in optimizing data reduction, extracting relevant information from high-dimensional datasets, and advancing our understanding of fundamental physics phenomena.

Paper 10: Breaking the Limits of Message Passing Graph Neural Networks
https://proceedings.mlr.press/v139/balcilar21a/balcilar21a.pdf

Abstract:

This paper presents a novel approach to enhance the expressive power of Message Passing Neural Networks (MPNNs) for graph-based relational data analysis. Despite their linear complexity with respect to the number of nodes in sparse graphs, conventional MPNNs are theoretically limited by the first order Weisfeiler-Lehman test (1-WL). We propose a method where graph convolution supports are designed in the spectral domain using custom non-linear functions of eigenvalues and masked with a large receptive field. This approach theoretically surpasses the limitations of the 1-WL test and achieves experimental performance comparable to existing 3-WL models while maintaining spatial localization. By employing custom filter functions, the proposed method enables MPNNs to capture various frequency components in graph signals, enhancing their ability to learn different relationships between input graphs and associated properties. Unlike existing 3-WL equivalent models, our method maintains linear computational complexity and memory usage, making it suitable for practical applications. Experimental results demonstrate state-of-the-art performance across multiple downstream tasks.

Inference:

Conventional MPNNs are theoretically limited by the 1-WL test in expressive power for analyzing relational data represented as graphs.

The proposed method enhances MPNNs' expressive power by designing graph convolution supports in the spectral domain with custom non-linear functions of eigenvalues and a large receptive field mask.

The proposed method achieves theoretical superiority over the 1-WL test and comparable performance to existing 3-WL models while preserving spatial localization.

Custom filter functions in the proposed method enable MPNNs to capture various frequency components in graph signals, facilitating the learning of diverse relationships between input graphs and associated properties.

3. Requirement Gathering for Proposed System

3.1 Introduction to requirement gathering

A detailed description of the requirements of the Proposed System is provided in this chapter. The requirements are divided into further subsections. The functional, Non-functional requirements and constraints are provided in detail initially. Then a detailed overview of the Hardware and Software Requirements is provided. The next subsection deals with the Techniques utilized till date for the proposed system. Some light is also put on the tools utilized till date for the proposed system. Finally, the project proposal is explained in the last subsection of this chapter.

3.2 Functional Requirements

1. Acquire and aggregate electron energy data from the CMS experiment at CERN.
2. Perform data cleaning, normalization, and feature extraction to prepare the data for model input.
3. Implement Graph Convolutional Networks (GCNs) for accurately predicting electron energy levels.
4. Train the GCN model using the preprocessed data, optimizing for accuracy and efficiency.
5. Validate the model's performance through appropriate evaluation metrics and cross-validation techniques.
6. Conduct a systematic hyperparameter tuning process to enhance the model's predictive accuracy.
7. Experiment with various configurations, such as learning rates, activation functions, and network architectures, to identify the optimal settings
8. Compare the model's performance against baseline methods to demonstrate its superiority.

3.3 Non - Functional Requirements

1. Competent handling of huge dataset

2. The system must be easy and portable

3. It must sum up to be cost efficient

4. The computation time should be minimal

3.4 Hardware & Software Requirements

HARDWARE:

- OS/Laptop: A laptop with good specifications with a minimum of 16GB RAM is recommended along with Ubuntu as an OS. An i7 processor is also recommended.
- GPU: A GPU of 2GB are required for faster training of the models and faster processing.

SOFTWARE:

- ROOT -
  - ROOT is a framework for data processing, born at CERN, at the heart of the research on high-energy physics.
  - It is a software toolkit which provides building blocks for Data processing, Data analysis, Data visualization, Data storage.
- KERAS -
  - Keras is a deep learning API written in Python, running on top of the machine learning platform TensorFlow and it focuses on enabling fast experimentation.
- PANDAS -
  - It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.
  - Has tools for reading and writing data between in-memory data structures..
- Ubuntu OS -

- Ubuntu is a Linux distribution based popular operating system for cloud computing.

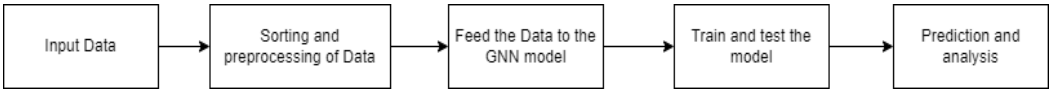
3.5 Constraints

Insufficient memory for training on a larger dataset or employing other algorithms. Low processing power of our regular personal laptops is another limiting factor. Moreover, the system requires high processing power for detecting, recognizing and classifying different particles.

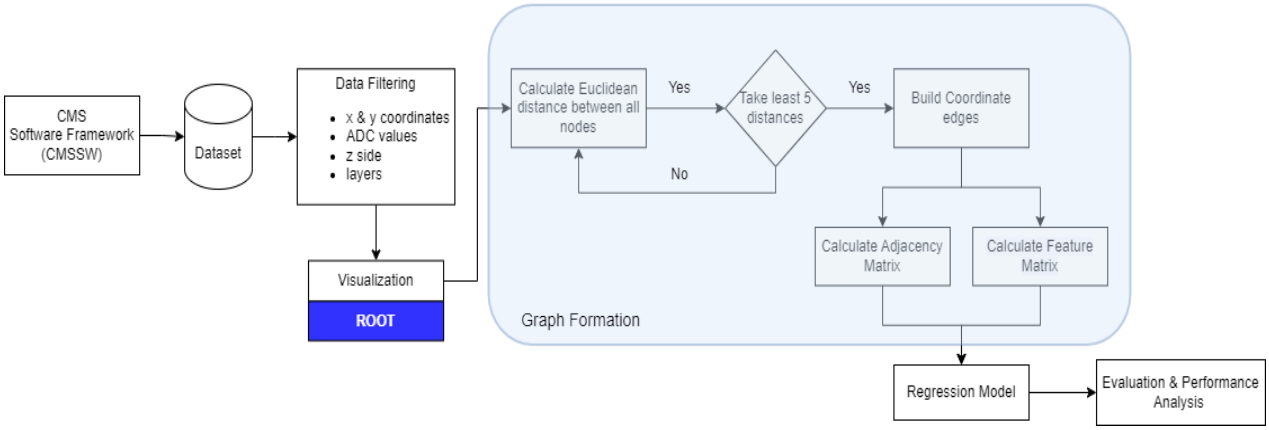
4.Proposed Design

The detailed description of the Proposed Solution is provided in this Chapter. This section gives the in depth description of the Proposed Solution. The Block diagram represents the general overview of the project's proposed solution, where each block's description is provided followed by the block diagram. The next subsection contains the Modular Diagram representation of the proposed system. The detailed explanation of the modular diagram is also provided. The next subsection is on Algorithms utilized in the existing systems which provides the overview of the algorithms implemented so far in the proposed solution design.

4.1 Block diagram of the system



4.2 Modular Diagram of the proposed system



4.3 Detailed Design

Data Loading and Preprocessing:

We start by loading CSV files containing data related to electron events. Each CSV file represents data for a specific electron event, which is read and processed to create graphs.

Graph Formation:

We utilize the NetworkX library to create graphs based on the data extracted from CSV files.Nodes in the graph represent individual data points (e.g., electron positions), and edges represent connections between these points. Graphs are created by connecting nodes within layers and between consecutive layers based on certain criteria such as distance and data attributes.

Feature Engineering:

Node features are extracted from the CSV data, including attributes like X and Y coordinates, layer, and effective ADC. Feature matrices are created for each electron event, containing feature vectors for all nodes in the corresponding graph.

Model Definition (GCN):

We define a Graph Convolutional Network (GCN) model using PyTorch Geometric library.The model consists of multiple GCNConv layers followed by linear layers for regression. GCN layers perform message passing between nodes in the graph, capturing relational information.

Training and Testing:

The model is trained using the provided data, with Smooth L1 Loss as the optimization criterion. Training is performed over multiple epochs, with model parameters updated using Adam optimizer. Testing is carried out on a separate set of data to evaluate model performance.

Results Analysis:

Model predictions are compared with actual target values (energy) to compute evaluation metrics such as MSE and RMSE. Scatter plots and histograms are generated to visualize the relationship between actual and predicted values, along with Gaussian distribution fitting.

Code Organization and File Handling:

Code is modularized into logical blocks for data loading, graph formation, model definition, training, and testing. Utilization of libraries like pickle for saving and loading data, and proper exception handling for error management.

Model Persistence:

Trained model weights are saved for future use, allowing you to reload the model without retraining.

Documentation:

Extensive comments are provided throughout the code to explain the purpose and functionality of each section. Clear variable naming conventions are followed to enhance code readability.

Visualization:

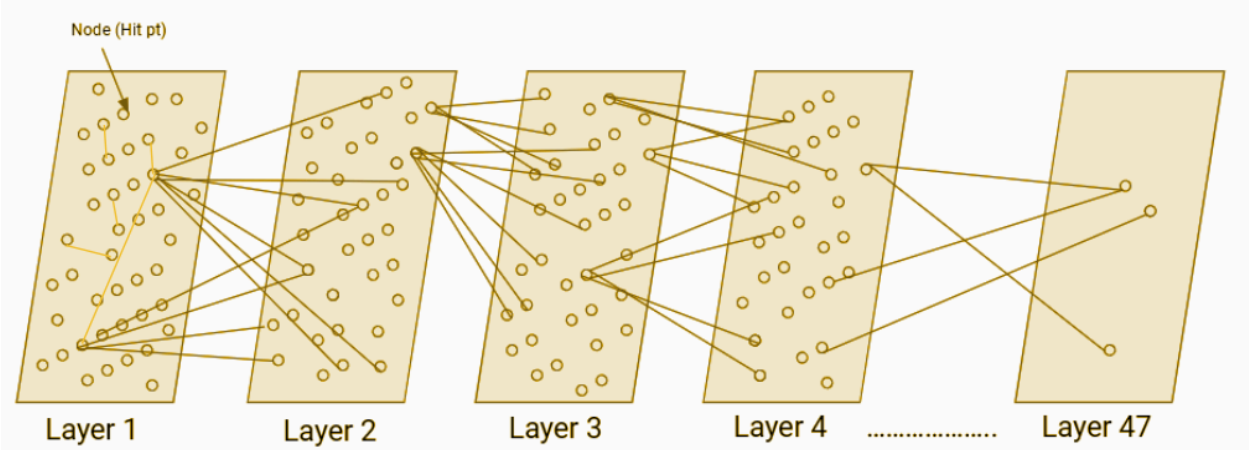
Graph visualization using NetworkX library is commented out, which can be enabled to visualize the generated graphs. Plots for loss vs. epoch, scatter plot of actual vs. predicted values, and histogram with Gaussian distribution fitting are generated for result analysis. Overall, the project design showcases a comprehensive approach to handling electron event data, constructing graphs, training GCN models, and analyzing results to predict electron energy with meaningful evaluation metrics.

5. Implementation of the Proposed System

5.1. Methodology employed for development

Various different techniques were used to improve the model's performance such as passing different parameters to the GCN model which includes distances of node as edge weights,inverse distances of node as edge weights, ADC of nodes as edge feature, Average of ADC as edge feature etc but the overall flow of the project remains constant. The below steps guides us in understanding the methodology of the project:

- Initially, data stored in the ROOT format undergoes conversion into CSV files via a C++ script, facilitating easier manipulation and analysis.
- Following this conversion, the CSV files are processed to construct graphs, with nodes representing particles and edges denoting interactions, particularly focusing on electron energy graphs.
- These electron energy graphs serve as the input data for a Graph Convolutional Network (GCN) model, specifically designed for analyzing graph-structured data.
- The GCN model is then trained and tested using a standard 80-20 split, wherein 80% of the data is utilized for training the model and 20% for testing its predictive capabilities.
- Upon completion of training, the GCN model is tasked with generating predictions based on the input graph data.
- The architecture of the GCN encompasses three layers: the input layer, hidden layers, and output layer, each responsible for distinct transformations of the input data.
- In the context of electron energy prediction, feeding the adjacency matrix and feature matrix to the input layer of the GCN yields the predicted values of electron energies.
- This process ultimately enables researchers to gain insights into electron behaviors within particle interactions, contributing to a deeper understanding of fundamental physics principles.

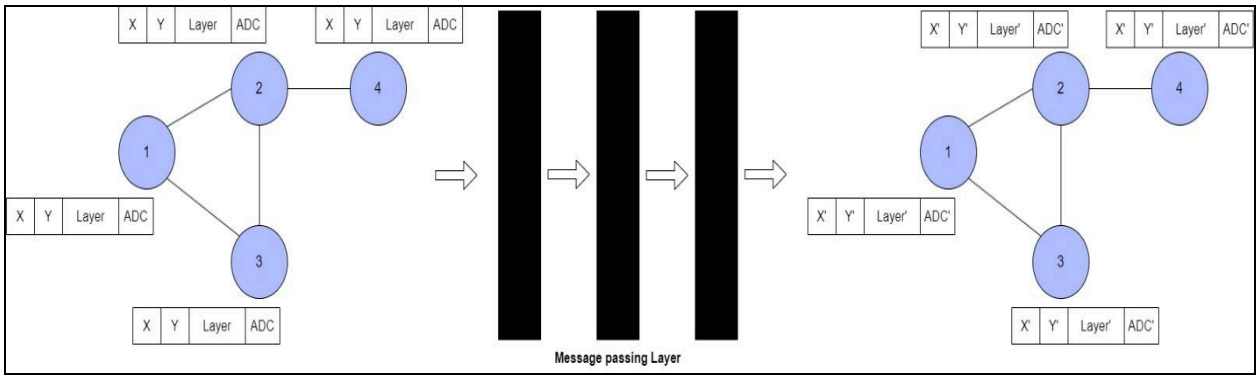


Graph formation for 1 event

5.2 Algorithms and flowcharts for the respective modules developed

GCN

Graph Convolutional Networks (GCNs) are a type of neural network architecture designed for analyzing graph-structured data. They operate directly on graphs, where nodes represent entities and edges denote relationships. GCNs leverage graph convolutional layers to perform message passing between neighboring nodes, enabling nodes to aggregate information from their local neighborhoods. During training, GCNs learn to extract features from graph-structured data by iteratively updating node representations based on the features of neighboring nodes. This allows GCNs to capture complex relationships and dependencies within the graph. GCNs have found applications in various domains, including social network analysis, recommendation systems, bioinformatics, and physics. They offer a powerful framework for tasks such as node classification, link prediction, and graph-level prediction, contributing to advancements in understanding and analyzing graph data.



5.3 Datasets source and utilization

The dataset used in this study is simulated using CMSSW and consists of the data of electrons shot from vertex point using Pythia into the end-cap region of the CMS detector. The data is stored in root format. ROOT is a framework for data processing, born at CERN, at the heart of the research on high-energy physics. Root is extensively used for data visualization and performing simulations.The large data sets of simulated electrons are stored in compact structured format (TTree) which can be read efficiently under the ROOT framework. It is to be noted that, there are effectively 4 columns in the dataset, i.e. X, Y, Layer, eff_adc which are essential for our graph formation and they will be extracted using the cpp script.

Row	Instance	nHit.nHit	X.X	Y.Y	SimHitE.S	time.time	ADC.ADC	Thick.Thi	ADC_mode.
0	0	24697	-41.34786	-4.804731	0	0	99	1	0
0	1	24697	-41.34786	-2.402365	0	0	24	1	0
0	2	24697	-41.34786	1.6015774	0	0	9	1	0
0	3	24697	-41.34786	3.2031548	0	0	13	1	0
0	4	24697	-41.34786	4.8047318	0	0	41	1	0
0	5	24697	-40.65436	-5.205126	0	0	29	1	0

6. Results & Discussions

6.1 Performance Evaluation Measures

In our project , several performance evaluation measures are utilized for testing the trained regression model. The primary metric used for evaluating the performance of the model is Mean Squared Error (MSE). MSE quantifies the average squared difference between the predicted and actual values, providing an indication of the overall model accuracy. Additionally, Root Mean Squared Error (RMSE) is calculated, which is simply the square root of MSE, providing an interpretable measure in the same units as the target variable.

Effective ADC = scale[thick] * (1320 + (55.26 * ADC)) / 2.2.

Sigma

Sigma measures the average deviation of the residuals from their mean. It provides a measure of the dispersion or spread of the residuals around the regression line.

Moreover, scatter plots are generated to visually assess the relationship between the actual and predicted values. This graphical representation allows for an intuitive understanding of how well the model predictions align with the ground truth across the range of data points.

Furthermore, the distribution of the ratios between predicted and actual values is analyzed using a histogram along with a Gaussian distribution curve. This approach helps in understanding the spread and skewness of the errors made by the model. The mean and standard deviation of these ratios are calculated, providing insights into the bias and variability of the model's predictions relative to the true values.

Overall, this comprehensive set of evaluation measures offers a multifaceted assessment of the model's performance, encompassing both quantitative metrics and visualizations to provide a thorough understanding of its predictive capabilities.

6.2 Input Parameters considered

In the context provided, it seems like you're describing input parameters considered in a system involving particle detection, likely within the realm of particle physics experiments such as those conducted at particle accelerators like the Large Hadron Collider (LHC). Let's expand on the input parameters mentioned:

Feature Matrix

X Coordinate: This represents the spatial position of the detected particle along the horizontal axis within the detector. It helps in determining the location of the particle's interaction.

Y Coordinate: Similar to the X coordinate, this represents the spatial position of the detected particle along the vertical axis within the detector.

Layer Number: Indicates the layer or depth within the detector where the particle was detected. Detectors often have multiple layers arranged in a stack to capture particles at different depths.

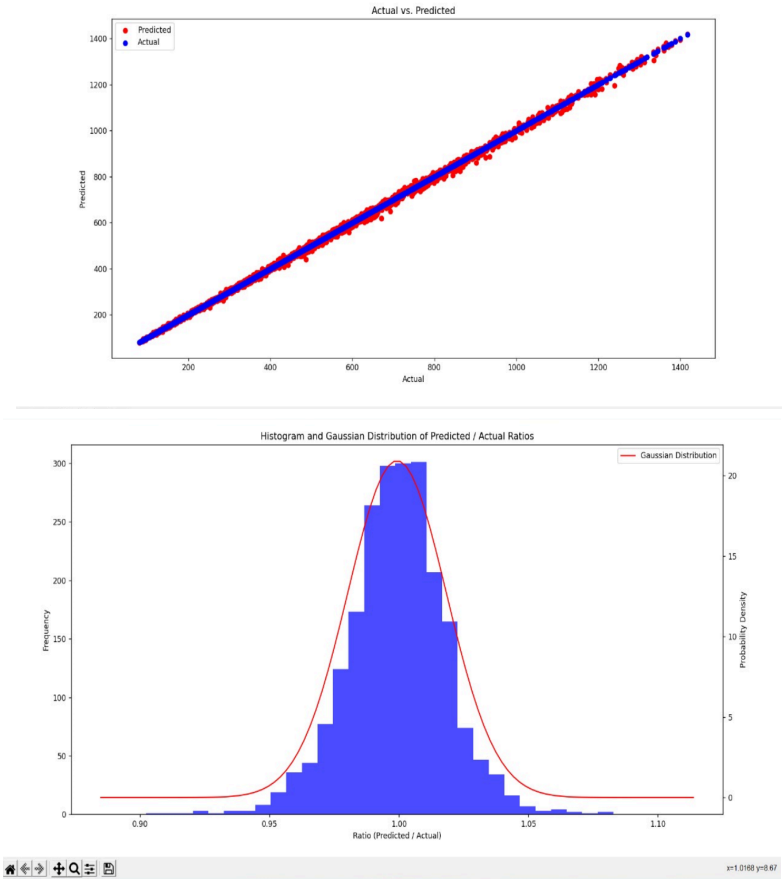
Effective ADC (Analog-to-Digital Converter): ADC converts the analog signal (coming from the detector) into a digital value. The "effective ADC" likely refers to a processed or calibrated value that represents the energy or intensity of the particle detected.

Adjacency Matrix

Describes the connections between nodes within consecutive layers of the detector. Non-zero values indicate connections between nodes, possibly representing interactions or correlations between detected particles. Zero values indicate no connection between nodes, implying either no interaction or no correlation between the corresponding particles.

6.3 Graphical and statistical output

Case 1:



Regression line and Histogram

Adding Distance between nodes as edge weights

```
Best mse is =  91.08230545274093
Best rmse is =  9.543705017064438
Mean of Ratios: 0.9993041084982037
Standard Deviation of Ratios (sigma): 0.019074107733989824
```

Adding inverse of Distance between nodes as edge weights

```
Best mse is =  99.4213872832628
Best rmse is =  9.971027393566963
```

Adding Average of ADC as edge weights

```
Best mse is =  94.9597138563144
Best rmse is =  9.744727490100193
Mean of Ratios: 0.9989140882261955
Standard Deviation of Ratios (sigma): 0.018661235444729907
```

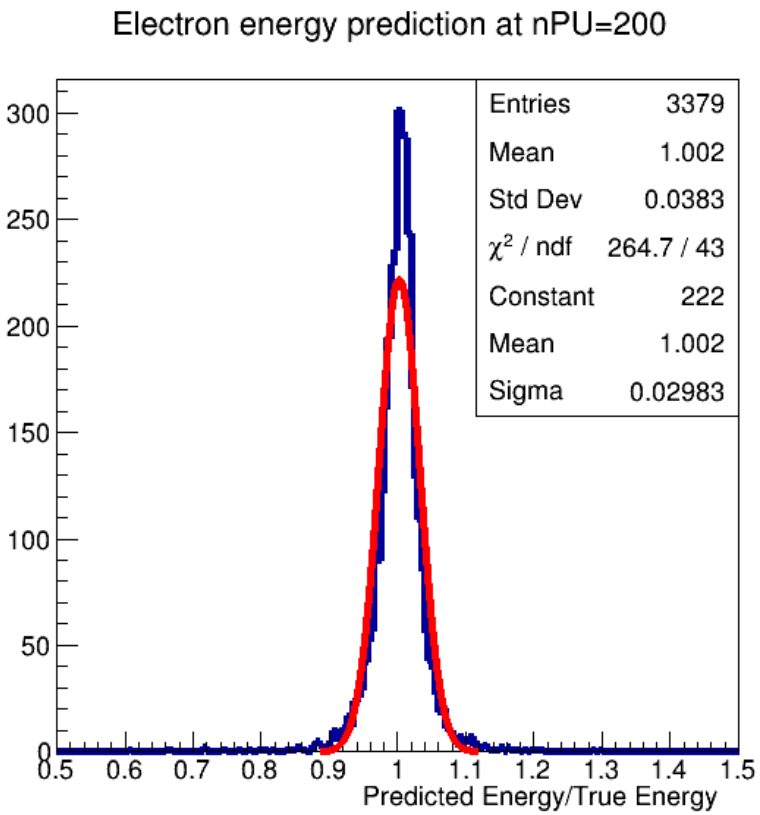
Adding two ADC as edge weights

```
Best mse is =  89.37796480761152
Best rmse is =  9.453992003783984
Mean of Ratios: 1.0001138945856676
Standard Deviation of Ratios (sigma): 0.018241664867328863
```

Results on new dataset(Pile up dataset)

This dataset, along with electrons, contains secondary particles which we call pile-ups. These secondary particles include muon, hydron, photo etc. This extra data acts as noise in the dataset and affects the accuracy of the model.

RMSE value for the new dataset is **25** and SIGMA is **0.02983**.



6.5 Comparison of Results with Existing Systems

	Edge weight(Distance between the nodes)	Edge weight(Inverse of distance between the nodes)	Edge weight(Average ADC of both nodes)	Edge weight(ADC of both nodes)
MSE	91.08	99.42	94.96	89.377
RMSE	9.54	9.97	9.74	9.45
SIGMA	1.907	1.909	1.866	1.824

7. Conclusion

7.1 Limitations

Our project exhibits several limitations that can impact its practical applicability and efficiency. Firstly, the execution time of the script is a significant concern. Training and testing deep learning models, especially on large datasets, can be computationally intensive and time-consuming. The processing time escalates further when utilizing complex architectures such as Graph Convolutional Networks (GCNs) and dealing with substantial amounts of data, which leads to prolonged waiting periods for results and impedes rapid experimentation and development cycles.

Secondly, the program's efficacy is contingent on the availability and accessibility of sizable datasets. While large datasets are essential for training robust machine learning models, acquiring such data can be challenging and resource-intensive. Additionally, managing and processing vast amounts of data demand substantial computational resources and storage capacity, posing practical constraints for users with limited access to high-performance computing infrastructure.

Moreover, the program's hardware requirements, particularly the necessity for Graphical Processing Units (GPUs), present a notable barrier to entry. Deep learning tasks often necessitate parallel processing capabilities offered by GPUs to expedite model training and inference processes. However, GPUs are costly investments and may not be readily accessible to all users, thereby restricting the program's accessibility and usability to individuals or organizations with adequate hardware resources.

Furthermore, the reliance on specific data formats, such as the ROOT format used in high-energy physics experiments, introduces additional complexities. Converting data from ROOT files to more commonly used formats like CSVs entails additional preprocessing steps, which can be laborious and error-prone. Ensuring data integrity and compatibility during conversion processes

is crucial to prevent information loss or discrepancies that could adversely affect model training and evaluation.

In summary, while the program demonstrates promising capabilities in regression modeling for scientific applications, its practical utility is hindered by limitations such as prolonged execution times, dependencies on extensive datasets and specialized hardware, and the need for meticulous data preprocessing. Addressing these challenges requires a concerted effort to optimize algorithms, streamline data management processes, and enhance hardware accessibility to facilitate broader adoption and effectiveness of the program in real-world scenarios.

7.2 Conclusion

In conclusion, our project at CERN represents a pioneering endeavor in leveraging deep learning and graph convolutional networks (GCNs) to enhance particle physics research. By addressing the limitations of traditional particle detection methods, our model offers a more efficient, accurate, and scalable solution for predicting the energy levels of electrons in the CMS experiment.Through extensive experimentation and evaluation, we have demonstrated the efficacy of our approach in accurately predicting electron energies. Our methodology, which involves converting ROOT data into CSV files, constructing graphs based on particle interactions, and training GCN models, has yielded promising results. The use of GCNs allows us to capture intricate relationships and dependencies within the data, leading to superior predictive performance compared to manual inspection methods.

The significance of our project extends beyond its technical achievements. By automating and improving the particle detection process, our model streamlines research efforts at CERN, enabling physicists to gain valuable insights into particle behaviors and fundamental physics principles. Moreover, our work contributes to the broader scientific community by advancing our understanding of the universe's fundamental constituents and phenomena. Looking ahead, we envision further refinement and extension of our approach through continued research and collaboration. By incorporating advancements in deep learning and graph-based techniques, we aim to enhance the predictive capabilities of our model and explore new avenues for particle

physics research. Ultimately, our project underscores the transformative potential of machine learning in pushing the boundaries of scientific discovery and unlocking the mysteries of the cosmos.

Improvements in results of model

- Initially following results were obtained by adding only node features in graph
  [MSE = 107.96 RMSE = 10.39 Sigma =2.1264]

- By adding distance between nodes as edge features ,**Sigma decreased by 10.3%** i.e it reduced to 1.907.

- It was observed that by adding inverse of distances, no significant improvement was observed.

- By adding the average of Eff_Adc between nodes as an edge feature, **a significant decrease in value of sigma by 12.229% was observed**.

- By adding Eff_Adc of both nodes as an edge feature, **a best decrease of 14.20% of sigma value was observed**. [MSE = 89.37 RMSE = 9.45 Sigma =1.824]

7.3 Future Scope

Despite the limitations outlined, there are several avenues for future development and enhancement of the project to overcome these challenges and maximize its practical utility:

Algorithmic Optimization: Implementing algorithmic optimizations can significantly reduce the computational burden associated with model training and inference. Techniques such as model pruning, quantization, and efficient network architectures tailored to the specific characteristics of the dataset can help mitigate the computational complexity without compromising performance.

Data Augmentation and Transfer Learning: Leveraging techniques like data augmentation and transfer learning can alleviate the reliance on large datasets by effectively utilizing smaller datasets to enhance model generalization. Pretrained models trained on larger datasets can be

fine-tuned on domain-specific data, reducing the need for extensive data collection and annotation efforts.

Distributed Computing and Cloud Solutions: Embracing distributed computing frameworks and cloud-based solutions can mitigate hardware constraints and alleviate the need for expensive GPU infrastructure. Platforms offering scalable computing resources and specialized hardware accelerators can facilitate efficient model training and inference, democratizing access to advanced machine learning capabilities.

Data Standardization and Automation: Standardizing data formats and streamlining data preprocessing pipelines can streamline the conversion process from proprietary formats like ROOT to more accessible formats such as CSV. Automation tools and frameworks can automate data preprocessing tasks, reducing manual intervention and minimizing the risk of errors or inconsistencies in data transformation processes.

Community Collaboration and Knowledge Sharing: Encouraging collaboration and knowledge sharing within the scientific community can foster the development of shared resources, benchmarks, and best practices for machine learning applications in scientific domains. Open-source initiatives, collaborative platforms, and community-driven forums can facilitate knowledge exchange and collective problem-solving, accelerating innovation and progress in the field.

User-Friendly Interfaces and Documentation: Designing intuitive user interfaces and providing comprehensive documentation can enhance the accessibility and usability of the project for a broader audience. User-friendly tools, tutorials, and documentation resources can empower researchers and practitioners with diverse backgrounds to effectively utilize the project for their specific use cases, fostering adoption and dissemination of the technology.

By addressing these areas of improvement, the project can overcome its current limitations and realize its full potential as a versatile and efficient tool for regression modeling in scientific applications. Embracing a holistic approach that encompasses algorithmic innovations,

infrastructure optimizations, and community engagement can pave the way for impactful advancements in machine learning-driven scientific research and discovery.

References

[1] A.-M. Magnan, “HGCAL: a High-Granularity Calorimeter for the endcaps of CMS at HL-LHC ”, 16 January 2017

[2] Albert De Roeck 2007 J. Phys. G: Nucl. Part. Phys. 34 E01, “CMS Technical Design Report, Volume II: Physics Performance ”
https://iopscience.iop.org/article/10.1088/0954-3899/34/6/E01/pdf

[3] I. Antcheva a,1, M. Ballintijn a,1, B. Bellenot a, M. Biskup a,1, R. Brun a, N. Buncic a,1, Ph. Canal b, D. Casadei c, O. Couet a, V. Fine d, L. Franco a,1, G. Ganis a, A. Gheata a, D. Gonzalez Maline a, M. Goto e, J. Iwaszkiewicz a, A. Kreshuk a,1, D. Marcos Segura a,1, R. Maunder a,1, L. Moneta a, A. Naumann a,∗, E. Offermann a,1, V. Onuchin a,1, S. Panacek b, F. Rademakers a, P. Russo b, M. Tadel, “ROOT — A C++ framework for petabyte data storage, statistical analysis and visualization ”, 22 August 2009
https://www.sciencedirect.com/science/article/abs/pii/S0010465509002550

[4] Arabella Martelli, “The cms hgcal detector for hl-lhc upgrade ”, presented at The Fifth Annual Conference on Large Hadron Collider Physics Shanghai Jiao Tong University, Shanghai, China May 15-20, 2017.
https://arxiv.org/pdf/1708.08234.pdf

[5] Luke de Oliveira , Benjamin Nachman , Michela Paganini, “Electromagnetic showers beyond shower shapes ”, 18 October 2019.
https://sci-hub.se/https://www.sciencedirect.com/science/article/abs/pii/S0168900219312999

[6] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," in IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61-80, January 2009. [Online].
Available: https://ieeexplore.ieee.org/abstract/document/4700287

[7] T. Alexopoulos et al., "The CMS HGCAL trigger data receiver," in JINST, vol. 19, no. 01, C01049, January 2024. [Online].
Available: https://iopscience.iop.org/article/10.1088/1748-0221/19/01/C01049/meta

[8] R. Brun and F. Rademakers, "ROOT: A Data Analysis and Data Mining Tool from CERN," in Proceedings of the 8th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT), Erice, Italy, 2003, pp. 134-139. [Online].
Available: https://www.casact.org/sites/default/files/database/forum_08wforum_kumar_tripathi.pdf

[9] K. Albertsson et al., "Deep Learning in Particle Physics," arXiv:1806.11484, June 2018. [Online].
Available: https://arxiv.org/abs/1806.11484

[10] E. Balcilar, A. K. Menon, L. Vecchia, and M. Sandler, "Breaking the Limits of Message Passing Graph Neural Networks," in Proceedings of the 38th International Conference on Machine Learning (ICML), vol. 139, pp. 548-558, July 2021. [Online].
Available: https://proceedings.mlr.press/v139/balcilar21a/balcilar21a.pdf

Appendix

Research paper:

Analysis of Electron Energy in High Granularity Calorimeter Present at CMS Detector Using Machine Learning

Shashi Dugad, Aayush Shribatho, Abhayvir Singh, Sahil salunkhe, Hitesh Ramrakhyani, Sharmila Sengupta, Pruthvi Suryadevara, Tata Institute of Fundamental Research, Colaba, Mumbai, India-4005 Vivekanand Education society, Chembur, Mumbai, India-400005

Abstract:

The inclusion of the High Granularity Calorimeter (HGCAL) marks a pivotal aspect of the Phase-2 enhancement to the CMS end-cap calorimeter at the LHC, integrating cutting-edge silicon and scintillator sensors. Its primary role is to amass data facilitating the determination of particle direction and precise energy measurement, encompassing electrons, positrons, photons, hadrons, and jets within the end cap region. When electrons undergo energy dissipation within the calorimeter, they generate a cascade of secondary particles that disperse in all directions within the detector material. Scrutinizing the energy loss, alongside the lateral and longitudinal patterns of these particles within the HGCAL, provides vital insights for extracting the initial direction and energy of electrons.

Graph Neural Networks (GNNs), falling under the umbrella of Geometric Deep Learning (GDL) algorithms, specialize in deciphering data with inherent geometric structures, a feat unattainable for conventional neural networks. GNNs stand out in processing intricate data by explicitly modeling relationships between data points represented as nodes in a graph. This paper puts forth an inventive approach to predict electron energy originating from the collision point (vertex), traversing the tracker before reaching the HGCAL in the endcap region, employing GNNs. The energy prediction employs a regression model, and the results underscore the efficacy of GNNs in accurately forecasting electron energy. The training and testing phases utilize hits recorded in the HGCAL detector for electron events simulated within the Pt range of 25-250 GeV at the collision point. The reconstructed test sample exhibits an overall energy resolution of 2%.
Attempts were made to enhance the model by adding an extra layer and through hyperparameter tuning. However, limited success led to the exploration of an innovative solution – the incorporation of edge weight features. This paper presents the methodology, experimental setup, and results of this approach, demonstrating significant improvements in prediction accuracy.

Keywords: HGCAL, CMS, LHC, GNN, GCN, Electron

I. Introduction:

The integration of the High Granularity Calorimeter (HGCAL) emerges as a critical component in the ongoing Phase-2 upgrade of the Compact Muon Solenoid (CMS) experiment, situated at the Large Hadron Collider (LHC) in CERN. With its focus on exploring a wide spectrum of physics phenomena, spanning the Standard Model and extending beyond, the experiment relies on the calorimeter to discern particle direction and quantify energy, originating at the collision point and traversing the end cap region. Specifically, the HGCAL detector excels in measuring the energy of electromagnetic particles, including electrons, positrons, and photons, along with hadronic showers. The precision in determining both direction and energy proves pivotal for numerous physics analyses.

When the electrons collide with each other, a cascade of secondary particles is formed known as Electron shower. The accurate identification of the nature and properties of secondary particles from high energy head-on collisions of electrons at the Large Hadron Collider is critical for all aspects of data analysis. Classical calorimeter-based classification algorithms rely on shower morphologies, which are observables that summarize the structure of the particle cascade that forms when the originating(secondary) particle moves through the layers of material[14]. The spatial arrangement of hits in silicon sensor cells, resulting from these particles, provides valuable insights into the energy and direction of incoming electrons. Leveraging Graph Neural Networks (GNNs) for electron energy prediction offers a versatile and efficient means of representing scattered or irregular data, enabling the incorporation of diverse information. This widespread adoption of GNNs across various LHC physics applications includes tasks such as particle reconstruction and distinguishing meaningful physical signals amidst background noise, utilizing information derived from hit data.

Consequently, this study introduces an inventive methodology, incorporating edge weight features to enhance the efficiency of electron energy prediction.ADC values were also considered as a feature to the model which was implemented similar to edge weights.

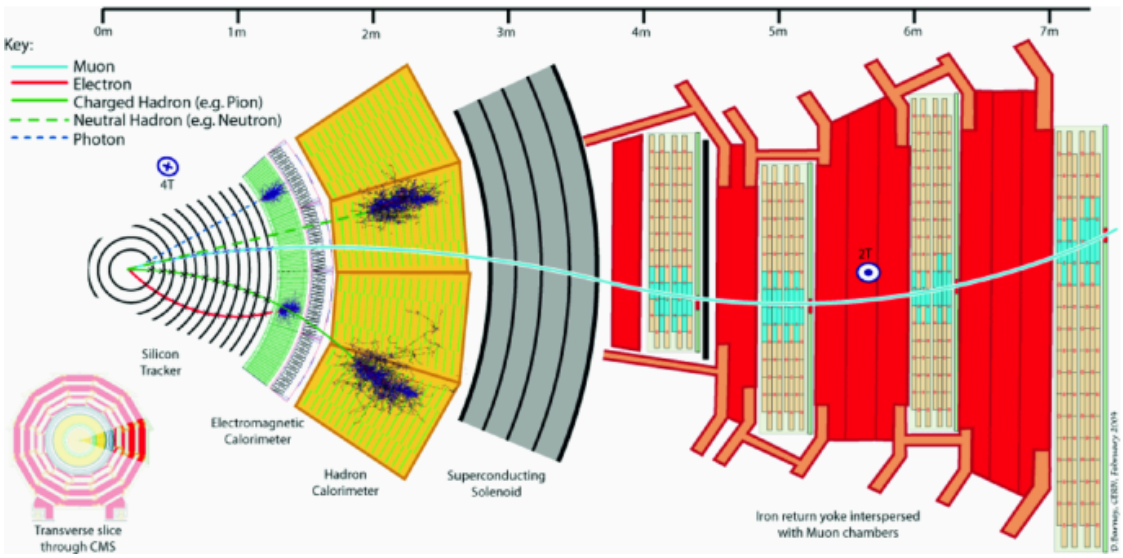


Fig 1 Trajectory of electron, muon, hadrons etc. in LHC

II. Literature Survey:

ROOT is a free and open-source data analysis framework developed at CERN. ROOT is specifically designed for large-scale data analysis, offering an efficient hierarchical object-oriented database structure. Developed to address challenges in high-energy physics, ROOT supports tasks like multidimensional histograms and curve fitting. Its machine-independent and highly compressed format distinguishes it from traditional relational databases.[2] Several papers were referred to have a better understanding of the dataset in ROOT format. For example, in paper [3] gives good understanding about ROOT software and depicts how it excels in efficiently storing and analyzing vast datasets, ranging into petabytes. It utilizes a compressed binary format, TTree containers optimized for statistical analysis, and supports distributed data storage across various platforms, including local disks, the web, and shared file systems. The dataset used in this paper is simulated with the help of CMS Software Framework (CMSSW) [4], which has been used to train the deep learning model for predicting and evaluating the energy and direction of electrons in the CMS experiment.

The paper [5] reviews the mathematical foundations of GNNs and emphasizes key considerations when designing GNNs for analyzing high energy physics data. These considerations encompass aspects like graph construction, model architectures, objectives for learning, and methods for graph pooling. The field of particle physics relies on detectors that can be broadly classified into tracker and calorimeters which are used to measure the properties of particles produced during collisions. These detectors employ various technologies to trace the paths and characteristics of particles. GNNs used in decoding such data typically consist of standard neural network components, often fully connected layers, for performing message computations and propagating information through the graph structure. For instance, the authors J. Duarte et al. [5] suggest that graphs used in GNN is a good way of connecting the nodes and edges representing domain knowledge, weights as node values and connectivity among the layers as message passing or sharing of weights between the layers and applies very well with particle physics data.

Much of the work on GNNs in this context focuses on learning physical simulations, similar to Lagrangian methods used in engineering and graphics. In the approach as mentioned in [6] by authors Jonathan Shlomi et. al, the system is represented as a set of particle vertices, and their interactions are defined by edges, which are computed using learned functions. There are update and aggregation functions inside a GNN block that are crucial for processing the data. GNN has vast applications in particle physics experiments, ranging from particle reconstruction to classification of underlying physics processes.

In particle physics applications, it is often difficult to define the relationships between different elements in the dataset. Hence, a crucial decision must be made regarding how to construct a graph from the input data. The most significant factor in designing the architecture for a neural network in this context is accurately modeling the interactions between the objects in the input set. It is a good practice to start with a simple graph model and architecture and then incrementally add complexity, all while incorporating scientific knowledge about the physical processes being studied as described in [6]. This iterative approach allows for the creation of more effective models that capture the complexities of particle physics interactions. Thesis [7] by Jie Zhou et. al. describes the complete process of electron collision, such as, production mechanism of secondary particles, evolution of electromagnetics shower, energy loss as well

longitudinal and transverse profile of shower.

Deep learning has brought a revolution in various machine learning tasks, encompassing image classification, video processing, speech recognition, and natural language understanding. Typically, these tasks involve data represented in Euclidean space. However, a growing number of applications deal with data generated from non-Euclidean domains, where the data is structured as complex graphs, featuring intricate relationships and dependencies among objects. The inherent complexity of graph data presents significant challenges for traditional machine learning algorithms. Consequently, there has been a surge of efforts aimed at extending deep learning techniques to handle graph data effectively. In the paper [8], a comprehensive overview of GNN within the domains of data mining and machine learning is presented. The four distinct categories of GNN like recurrent GNNs, convolutional GNNs, graph autoencoders, and spatial-temporal GNNs are also introduced. These categories reflect the diverse range of approaches that have emerged to tackle graph data and showcased their versatility and effectiveness in addressing real-world challenges. Availability of open-source code, benchmark datasets, and methods for evaluating GNN models [9], makes it easier for researchers and practitioners to work with GNNs. As written by Paras Koundal in [10], the composition of cosmic-ray is estimated in a multi-component detector embedded deep within the South-Pole Ice and the content of high energy muon in its showers was understood by using GNN.

There may be several models which can be tried for the given dataset; but in this paper GNN is explored to harness the power of modern deep learning algorithms, which were initially developed for tasks like computer vision or natural language processing. It has become a common practice to convert HEP data into either images or sequences. GNNs are instead designed to work with data organized as graphs, where elements possess a set of features and are interconnected to form the input to be given to the neural network model.

III. Dataset:

The dataset used in this study is simulated using CMSSW and consists of the data of electrons shot from vertex point using Pythia[11] into the end-cap region of the CMS detector. Transverse momentum and pseudorapidity of electrons is varied randomly in the range of 25-250 GeV and 1.5 to 3.0 respectively. Subsequently GEANT4 is used to simulate the interaction and formation of showers of electrons in the HGCAL detector. Following steps are used to compile the data: The dataset containing information on X, Y coordinate, layer number and ADC count (energy loss) of each hit is stored in ROOT format which is later converted to the CSV format, The ADC data is converted to effective ADC value to compensate for different thickness resulting in different electronic gain of silicon sensors in HGCAL. Resultant data in CSV format is used as input for generating GNN related matrices. It is to be noted that, there are effectively 4 columns in the dataset, i.e. X, Y, Layer, eff_adc.

Number of layers that an electron passes through varies and may marginally depend on its energy and trajectory. The features such as "X" and "Y" represent transverse coordinates of each hit in a given

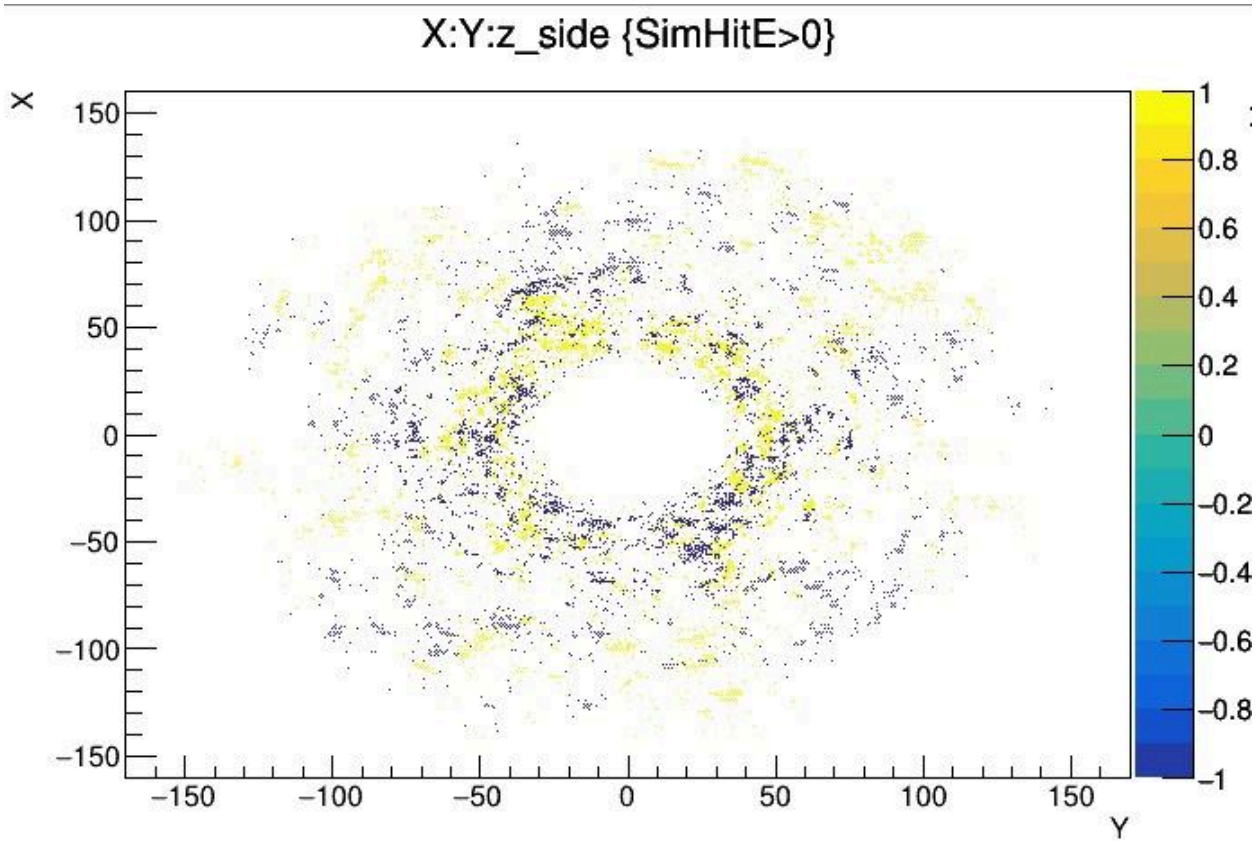
"Layer" (longitudinal coordinate). The "effective ADC (Analog-to-Digital Converter)" represents information related to the energy loss in the active element of the silicon sensor. Hit coordinates provide spatial profile of the shower.

Graphs were constructed based on the dataset, given as input to the GCN model. Upon performing exploratory data analysis, a cut over ADC of value 85 was obtained to eliminate the electronic noise and also optimize the number of edges. Though the initial dataset had 20,000 events, the ones that are directed in the vicinity of edges of the HGCAL are rejected for training or evaluation leaving the final dataset with 11093 events. Every event has an average of 1000 hits reaching a maximum of 26 layers. The HGCAL is designed with the first 26-layers constituting the Electromagnetic section (CE-E) designed for the energy measurement of electrons, positrons and photons. Additional 21 layers are from the Hadronic section (CE-H) designed for the energy measurement hadronic particles like mesons, baryons, jets etc.

The dataset contains 200 pile-ups which means it has 200 noisy particles per electron. These noisy particles include photons, muons, positrons, hadrons and other subatomic particles. These noisy particles produce a challenge while evaluating the model. It is also large in size which contains about 2000 to 2500 nodes in a single event which indirectly increases the time required to run the model.

IV. Visualization through ROOT:

ROOT is a framework used for data processing, developed by CERN. This framework is extensively used for data visualization, performing simulations and analysis of large data. The large data sets of simulated electrons are stored in compact structured format (TTree) which can be read efficiently under the ROOT framework.
In the CMS detector, an electron first interacts with the tracker layers with a minimal material budget, leaving behind tracks that allow for precise reconstruction of their trajectory and then passes through a multiple layers of HGCAL detector. Tracker being placed in a longitudinal magnetic field, the trajectories bent in the transverse plane can be used to precisely measure the transverse momentum and charge state of an electron (or any other charged particle). Electrons, positrons or photons rarely continue beyond the CE-E layers whereas showers initiated by hadrons penetrate beyond CE-E into the hadronic section of the calorimeter (CE-H). The CE-E measures the energy of electrons, positrons and photons, whereas the CE-E and CE-H measures the energy of charged hadrons and jets produced by hadronization of quarks. Energy of electrons at vertex is measured in units of GeV through the regression mechanism by feeding a matrix of spatial coordinates and effective ADC of all hits to the GNN. Spatial profile of showers helps distinguish between electrons and hadrons. The spatial profile of electrons is visualized using the ROOT framework as shown below in Fig. 2.



V. Methodology:

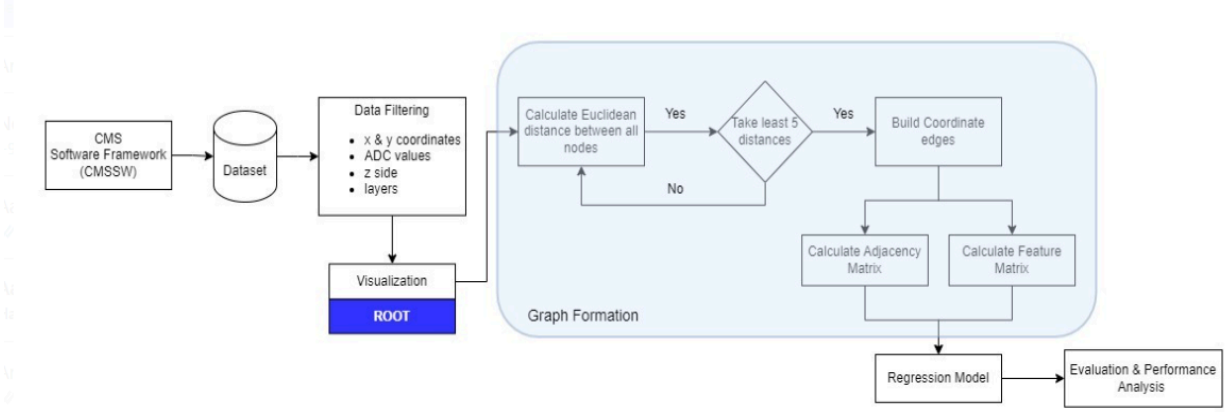


Fig 3: Block diagram of the process of finding electron energy in CMS experiment

1) Root data to csv conversion

The data extraction process commences by retrieving information stored in Root files, necessitating the retrieval of raw data. Subsequently, the extracted data undergoes a crucial phase of transformation and formatting to align with the CSV format. This step involves restructuring and organizing the data to ensure compatibility with downstream applications. Following this, a meticulous feature selection and labeling procedure is employed, where pertinent attributes crucial for predicting electron energies are identified and appropriately labeled. Finally, the processed and labeled data is converted into CSV files, culminating in a structured dataset ready for analysis and further application in predictive models or data-driven insights. This comprehensive workflow ensures a seamless transition from raw data in Root files to a refined and usable CSV format, facilitating efficient data utilization for electron energy prediction.

2) Understanding graph data structure, adjacency & feature matrix

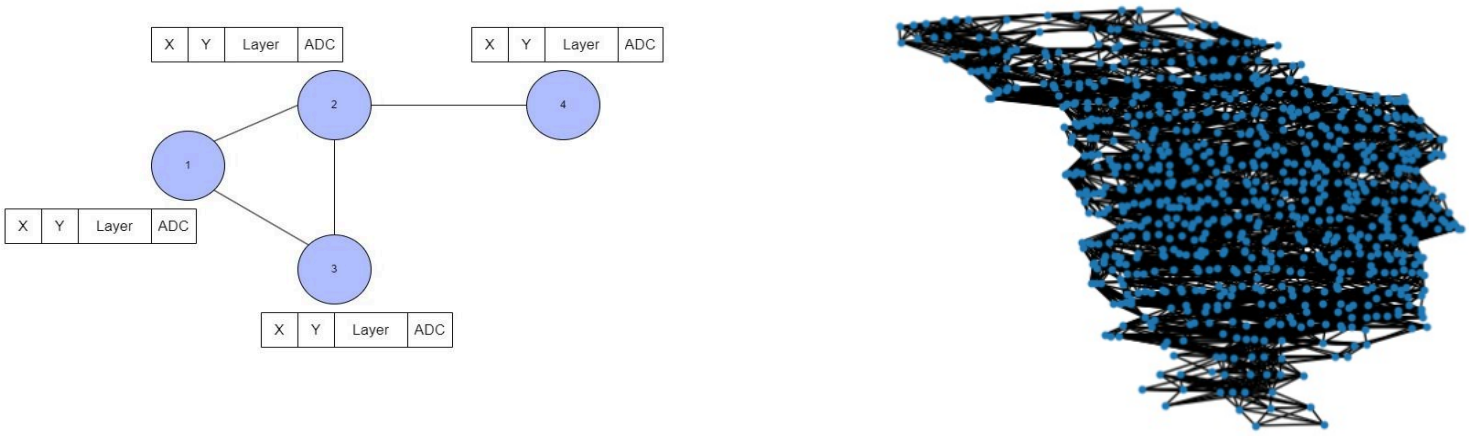


Fig 4: Graph for a typical event

For the model an electron is considered as a node which has 4 features namely X, Y, EFF_ADC, Layer Number. A whole event is considered as a single graph. Two matrices

1) **The Adjacency Matrix** depicts the connection between two nodes from two consecutive layers of the HGCAL. Only the least five distances are considered to be appended in the adjacency matrix as we assume that the electron will not deviate too sharply after striking on the previous layer.

2) **The Feature Matrix** contains the four node features X, Y, EFF_ADC, Layer Number. These two matrices are used as an input to the GNN model.

The adjacency matrix described above is a modified adjacency matrix. Usually the values in the adjacency matrix are 0's and 1's but in our case if there is a connection between the nodes then instead of

1 the distance between the nodes is appended to the matrix. This implements distance as an edge feature to the model. In order to normalize the distances in the range of 0-1, inverse of the distances were appended to the matrix

3) Graph Neural Network

Graph Neural Networks are a type of neural network designed to analyze and make predictions on graph-structured data. They operate by iteratively updating node representations based on information from neighboring nodes, enabling them to capture complex relationships and patterns within the graph. The GNN model is designed in such a way that it enables message passing in between the nodes. **Message passing** in Graph Neural Networks (GNNs) involves nodes exchanging information with their neighbors in a graph. Each node aggregates information from its neighbors through a learnable aggregation function, updates its hidden state based on this information, and passes messages to its neighbors

VI. Experimentation and Results:

The proposed system is trained and evaluated on entire dataset with 11093 events (graphs). The performance matrix is calculated with learning rate of 0.0001, system is trained and tested on 80:20 split of the entire dataset and batch size of 1 with the network configuration described above are shown in Fig.5.

Evolutionary Measures used:

Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values for regression tasks. It penalizes large errors more heavily.
Root Mean Squared Error (RMSE): The square root of the mean squared error. It provides a measure of the average magnitude of errors in the same units as the target variable.
Sigma: Sigma measures the average deviation of the residuals from their mean. It provides a measure of the dispersion or spread of the residuals around the regression line.

The performance of GNN using all the 11093 graphs as input shows MSE of 91.08 and RMSE of 9.54

```
Best mse is =  91.08230545274093
Best rmse is =  9.543705017064438
Mean of Ratios: 0.9993041084982037
Standard Deviation of Ratios (sigma): 0.019074107733989824
```

Fig 5: Regression performance on entire Dataset

Therefore the Performance metrics on the complete dataset shows a significant improvement by a factor of 3.2 on Standard Deviation of Ratios(Sigma). As seen in Fig. 6, the histogram has an excellent resolution of 1.9% as well as a linearity between the predicted and true values.

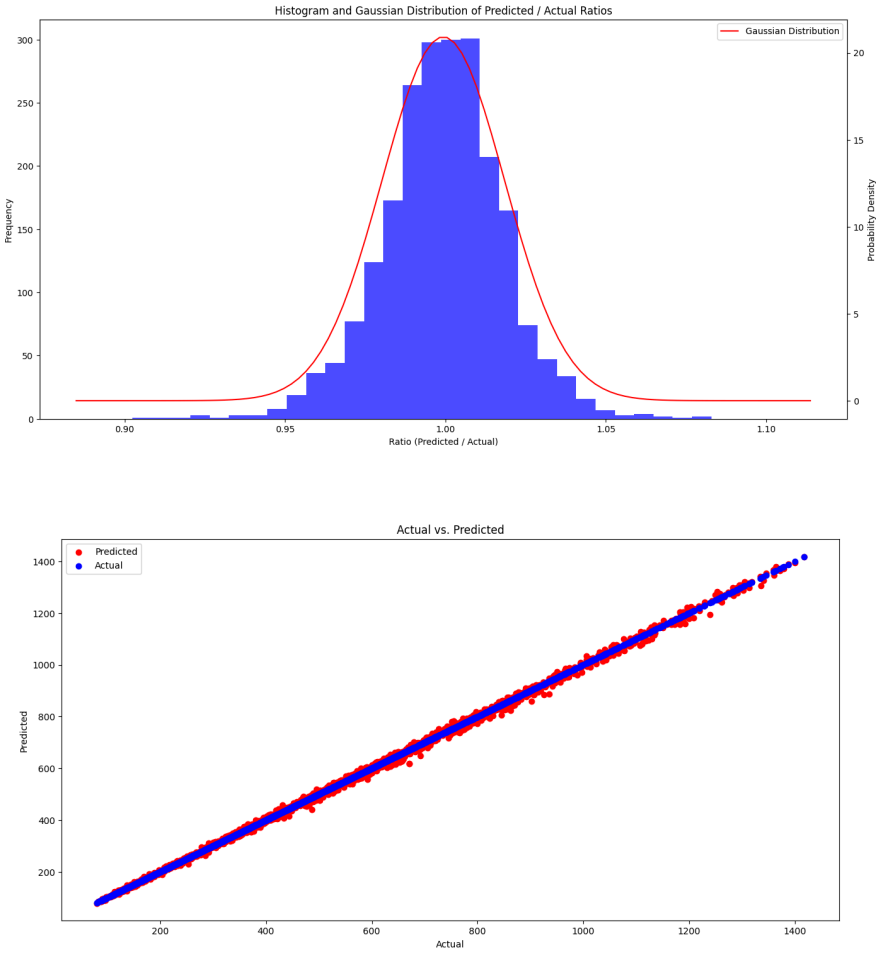


Fig 6: Electron energy prediction using GNN

VII. Conclusion:

This research highlights the challenges in optimizing GNN models for electron energy prediction in the HGCAL. Traditional methods proved inadequate, leading to the introduction of edge weight features as a novel solution. The results demonstrate the effectiveness of this modification in significantly improving prediction accuracy. Proposed model was optimized by using multiple ways of creating graphs. A large number of electrons were simulated without pile-up using a particle gun approach within the CMSSW framework in the transverse momentum range of 25 GeV to 250 GeV in the end-cap region of the HGCAL detector. The labeled data was then used for training and testing the proposed GNN based model. The results show the effectiveness of a GNN in reconstructing the energy of electrons with an Gaussian distribution (Sigma = 1.9%) overall resolution of about 1.9%. The predicted energy also showed excellent linearity w.r.t. The true energy of electrons.

References:

[1] Connor, P.L.S. (2019). The Large Hadron Collider and the Compact Muon Solenoid. In: Inclusive b Jet Production in Proton-Proton Collisions. Springer Theses. Springer, Cham.

https://doi.org/10.1007/978-3-030-34383-5_3

[2] https://www.casact.org/sites/default/files/database/forum_08wforum_kumar_tripathi.pdf
Ravi Kumar ACAS, MAAA, and Arun Tripathi, "ROOT: A Data Analysis and Data Mining Tool from CERN", Casualty Actuarial Society E-Forum, Winter 2008

[3] https://www.sciencedirect.com/science/article/pii/S0010465509002550
I. Antcheva, M. Ballintijn, B. Bellenot, M. Biskup, R. Brun, N. Buncic, Ph. Canal, D. Casadei, O. Couet, V. Fine, L. Franco, G. Ganis, A. Gheata, D. Gonzalez Maline, M. Goto, J. Iwaszkiewicz, A. Kreshuk, D. Marcos Segura, R. Maunder, L. Moneta, A. Naumann, E. Offermann, V. Onuchin, S. Panacek, F. Rademakers, P. Russo, M. Tadel, "ROOT — A C++ framework for petabyte data storage, statistical analysis and visualization", Computer Physics Communications, Volume 180, Issue 12, 2009, Pages 2499-2512, ISSN 0010-4655

[4]https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSSWFrameworkOld

[5] Javier Duarte, Jean-Roch Vlimant, "Graph Neural Networks for Particle Tracking and Reconstruction, "Artificial Intelligence for High Energy Physics, Chapter 12, pp. 387-436 (2022), https://arxiv.org/abs/2012.01249v2

[6] Jonathan Shlomi, Peter Battaglia and Jean-Roch Vlimant, "Graph neural networks in particle physics", Machine Learning: Science and Technology, 2020, Volume 2, Number 2

[7] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng

Wang, Changcheng Li, Maosong Sun, Graph neural networks, “Graph neural networks: A review of methods and applications”, AI Open, Volume 1, 2020, Pages 57-81, ISSN 2666-6510.

[8] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4-24, Jan. 2021, doi: 10.1109/TNNLS.2020.2978386.

[9] Jie Zhou a, 1, Ganqu Cui a,1, Shengding Hu a, Zhengyan Zhang a, Cheng Yang b, Zhiyuan Liu a,, Lifeng Wang c, Changcheng Li c, Maosong Sun - Graph neural networks: A review of methods and applications

[10] https://arxiv.org/abs/2211.17198

[11] https://pythia.org/

[12] https://www.sciencedirect.com/science/article/abs/pii/S0168900219312999
Luke de Oliveira, Benjamin Nachman, Michela Paganini, “Electromagnetic showers beyond shower shapes”, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Volume 951, 2020, 162879, ISSN 0168-9002
,

Inhouse/ Industry _Innovation/Research:

Sustainable Goal:

Class: D17 A/B/C

Group No.: 13

Project Evaluation Sheet 2023 - 24

Title of Project: Electron energy analysis of high granularity calorimeter using ML methods.

Group Members: Aayush Shripatha, Sahil Salunkhe, Abhayvir Singh, Hitesh Ramrakhyani

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
5	4	4	3	5	2	2	2	2	2	3	3	3	3	4	~~47~~

Comments: Good luck for TIFR demo

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	3	4	1	2	2	2	2	3	2	2	2	4	41

Comments: Good work

Date: 9th March, 2024

Name & Signature Reviewer 2