HYBRID DEEP LEARNING APPROACHES FOR CYBER THREAT DETECTION IN NETWORK AND SOCIAL MEDIA

A Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

MASTER OF TECHNOLOGY

in

DATA SCIENCE

by

VINAY RANA

(24/DSC/13)

Under the Supervision of

Dr. Rahul

Assistant Professor, Department of Software Engineering

Delhi Technological University, Delhi, India



Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2026

ACKNOWLEDGEMENT

It is with sincere appreciation that I thank the Almighty for constantly serving as the source of my strength, resilience and guidance during my research. Indeed, such a faith has given me a lot of enlightenment and motivation in the whole process of undertaking my research.

A heartfelt thanks to my supervisor, **Dr. Rahul**, Department of Software Engineering, Delhi Technological University for providing constant guidance, critical feedback, and encouragement during the course of my work. Under his guidance, the technical and research work of the project was undertaken.

I express my warm love and deepest gratitude to my parents their unconditional love, encouragement, and constant guidance. Without their support and motivating presence, it would not have been possible for me to achieve my goals. Their unwavering faith in me has been my greatest strength throughout this journey.

I would also like to express my sincere appreciation to my friends for their companionship and encouragement, which brought balance and positivity during the challenging phases of this journey. I am equally thankful to my colleagues and peers for their support and academic camaraderie. This thesis is the culmination of not just my efforts but the collective support, love, and encouragement of all these remarkable individuals. I forever remain grateful for the lasting impact they have had on my academic, professional, and personal life.

Vinay Rana
24/DSC/13



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

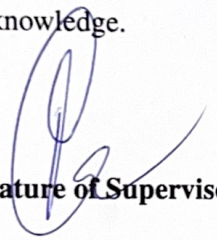
I Vinay Rana hereby certify that the work which is being presented in the thesis entitled Hybrid Deep Learning Approaches for Cyber Threat Detection in Network and Social Media in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Data Science, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from August 2024 to May 2026 under the supervision of Dr. Rahul.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

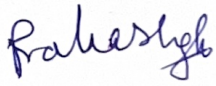


Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.



Signature of Supervisor(s)



Signature of External Examiner



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **Vinay Rana** (24/DSC/13) has carried out their research work presented in this thesis entitled **"Hybrid Deep Learning Approaches for Cyber Threat Detection in Network and Social Media"** for the award of **Master of Technology** from Department of Software Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Rahul
Assistant Professor
Dept. of Software Engineering, DTU

Place: Delhi
Date: 7/7/26

Deep Learning Approaches for Multimodal Human Activity Recognition

Vinay Rana

ABSTRACT

The rapid proliferation of digital platforms has given rise to two critical and structurally distinct categories of cyber threat: *cyberbullying* embedded in multimodal social-media content, and *network intrusion attacks* manifested in enterprise-level packet traffic. Existing detection systems address these threats in isolation and are characterised by a reliance on unimodal inputs, limited cross-modal reasoning, poor generalisation to class-imbalanced data, and an inability to capture the simultaneous spatial and temporal dependencies inherent in real-world attack behaviour.

For **Cyberbullying Detection**, a Hybrid Multimodal Image-Text Classification framework is proposed, evaluated on the Facebook Hateful Meme Dataset comprising 10,000 meme samples. Meme images are processed by a Dense Residual High-Resolution Network (DR-HRNet), which maintains high-resolution spatial feature maps throughout its four-stage parallel branch architecture and enriches them via a Dense Residual Convolutional Block (DRCB) that combines dense connectivity with residual learning. Meme captions are concurrently encoded by Multi-Scale DeBERTa (MC-DeBERTa), an enhanced transformer that exploits disentangled self-attention with relative position embeddings and stacks a Multi-Scale Dilated Convolutional Block (MDCB) atop the encoder to capture linguistic patterns at multiple granularities. The extracted visual and textual feature vectors are fused via Multimodal Compact Bilinear (MCB) Pooling, which models higher-order cross-modal interactions in a computationally efficient manner. The fused representation is classified by an Adaptive EfficientNetV2 backbone augmented with a Transformer Spatial Pooling (TSP) module that captures global contextual relationships before the final softmax decision layer. The proposed framework achieves a test accuracy of **99.5%** and an F1-score of **99.30%** on the Bully and Non-Bully classes respectively, with a macro-average AUC of **0.995**, significantly outperforming unimodal and conventional multimodal baselines.

For **Network Intrusion Detection**, a Graph-Based Hybrid GAT-BiLSTM framework is proposed, evaluated on the CICIDS2017 benchmark comprising approximately 2.83 million labelled network flows across five consolidated traffic categories (BENIGN, DDoS, PortScan, DoS, and Other). Each network flow is modelled as a node in a cosine-similarity k-Nearest Neighbour (kNN) graph with $k = 10$ and an adaptive dis-

tance threshold, enabling the structural relationships between flows to be explicitly represented. Two stacked multi-head Graph Attention Network (GAT) layers with batch normalisation, dropout, and residual connections learn spatial inter-flow topology by dynamically assigning differential attention weights to neighbouring nodes. The resulting node embeddings are processed by a Bidirectional Long Short-Term Memory (BiLSTM) network that captures long-range temporal dependencies in both forward and backward directions, modelling the sequential behavioural evolution of network traffic. A two-layer Multi-Layer Perceptron (MLP) head with a class-frequency weighted cross-entropy loss function performs the final five-class classification, counteracting the extreme class imbalance that is mitigated further through controlled random undersampling and stratified node splitting. Training employs the AdamW optimiser with ReduceLROnPlateau scheduling and early stopping. The proposed framework achieves a test accuracy of **94.20%**, precision of **94.53%**, recall of **94.20%**, and F1-score of **94.20%**, with a macro-average ROC-AUC of **0.997**, outperforming standalone GNN, CNN, MLP, and LSTM baselines on all metrics.

Collectively, this thesis demonstrates that hybrid deep learning architectures designed with specialised feature extractors, learned fusion or graph-aggregation mechanisms, and imbalance-aware training strategies provide a principled and effective approach to multi-domain cyber threat detection. The two proposed methodologies advance the state of the art in their respective domains and establish a generalised design philosophy for future cross-domain intelligent cybersecurity systems.

Keywords: Cyberbullying Detection, Dense Residual High-Resolution Network, Multi-Scale DeBERTa, Multimodal Compact Bilinear Pooling, Adaptive EfficientNetV2, Network Intrusion Detection, Graph Attention Network, Bidirectional LSTM, Class Imbalance.

LIST OF PUBLICATIONS

International Conferences

1. Rana, V. and Rahul, 2026. Hybrid Multimodal Image-Text Classification for Cyberbullying Detection using DR-HRNet and MC-DeBERTa with Adaptive EfficientNetV2. In *Proceedings of the 7th International Conference on Data Analytics & Management (ICDAM-2026)*. **(Scopus-Indexed, Accepted for Presentation, Scheduled during 12–14 June 2026)**

2. Rana, V. and Rahul, 2026. Graph-Based Hybrid GAT-BiLSTM Framework for Robust Intrusion Detection in Imbalanced Network Traffic. In *Proceedings of the 7th International Conference on Data Analytics & Management (ICDAM-2026)*. **(Scopus-Indexed, Accepted for Presentation, Scheduled during 12–14 June 2026)**

TABLE OF CONTENTS

Acknowledgement **i**

Candidate's Declaration **ii**

Certificate **iii**

Abstract **iv**

List of Publications **vi**

Table of Contents **vii**

List of Table[s] **x**

List of Figure[s] **xi**

List of Symbols, Abbreviations and Nomenclature **xii**

1 Introduction **1**

1.1 Background and Motivation 1

1.2 The Rise of Cyber Threats in the Modern Times 2

1.2.1 Cyberbullying as a Social Cyber Threat 2

1.2.2 Network Intrusion as an Infrastructure Cyber Threat 3

1.3 Problem Statement 4

1.4 Research Objectives 4

1.5 Key Contributions of the Work 5

1.6 Thesis Organisation 6

2 Literature Review **8**

2.1 Overview 8

2.2 Cyberbullying Detection 9

2.2.1 Classical Machine Learning Approaches 9

2.2.2 CNN and RNN Based Methods 10

2.2.3 Transformer-Based Methods 10

2.2.4 Multimodal Fusion Approaches 10

2.2.5 Recent Works (2024–2025) 11

2.3 Network Intrusion Detection Systems (NIDS) 12

2.3.1 Signature-Based vs Anomaly-Based IDS 12

2.3.2 Machine Learning and Deep Learning Approaches 13

2.3.3 Graph Neural Network Approaches 14
2.3.4 Encrypted Traffic, Flow-Level Detection, and Class Imbalance 14
2.4 Hybrid Deep Learning in Cybersecurity 15
2.5 Identified Research Gaps 16

3 Datasets and Preprocessing 18
3.1 Introduction 18
3.2 Cyberbullying Detection Dataset 18
3.2.1 Facebook Hateful Meme Dataset 18
3.2.2 Dataset Statistics and Split 19
3.3 Network Intrusion Detection Dataset 19
3.3.1 CICIDS2017 Dataset 19
3.3.2 Feature Groups and Label Distribution 20
3.4 Preprocessing Pipeline for Cyberbullying Detection 21
3.4.1 Image Preprocessing 21
3.4.2 Text Preprocessing 22
3.5 Preprocessing Pipeline for Network Intrusion Detection 23
3.5.1 Column Standardisation and Missing Value Handling 23
3.5.2 Removal of Non-Predictive Identifier Columns 24
3.5.3 Attack Label Cleaning, Grouping and Encoding 24
3.5.4 Feature Normalisation 24
3.5.5 Class Balancing via Random Undersampling 25

4 Proposed Methodologies 26
4.1 Methodology 1 - Multimodal Cyberbullying Detection 26
4.1.1 Framework Architecture Overview 26
4.1.2 Image Feature Extraction using DR-HRNet 28
4.1.3 Text Feature Extraction using MC-DeBERTa 29
4.1.4 Multimodal Feature Fusion using MCB Pooling 30
4.1.5 Classification using Adaptive EfficientNetV2 31
4.2 Methodology 2 - Graph-Based Hybrid Network Intrusion Detection . 32
4.2.1 Framework Architecture Overview 32
4.2.2 Graph Construction using k-Nearest Neighbours 32
4.2.3 Spatial Feature Learning using GAT 33
4.2.4 Temporal Feature Learning using BiLSTM 34
4.2.5 Classification using MLP 35
4.2.6 Training Strategy 35
4.3 Summary 36

5 Results and Discussion **37**
5.1 Implementation Details 37
5.1.1 Training Configuration and Hyperparameters 37
5.2 Results and Discussion – Cyberbullying Detection 39
5.2.1 Training and Testing Accuracy and Loss 39
5.2.2 Confusion Matrix Analysis 39
5.2.3 ROC Curve and AUC Analysis 41
5.2.4 Performance Comparison Metrics 41
5.2.5 Discussion of Results 44
5.3 Results and Discussion - Network Intrusion Detection 45
5.3.1 Confusion Matrix Analysis 45
5.3.2 ROC Curve and AUC Analysis 47
5.3.3 Performance Metrics Evaluation 48
5.3.4 Comparison with Baseline Models 49
5.3.5 Discussion of Results 52
5.4 Comparative Analysis Across Domains 53
5.4.1 Comparative Performance Analysis 53
5.4.2 Deep Learning Effectiveness Across Domains 53
5.4.3 Challenges in Multimodal and Network Data 54
5.4.4 Strengths and Limitations of the Proposed Frameworks 54
5.5 Summary of Key Results 55

6 Conclusion, Future Work, and Social Impact **56**
6.1 Conclusion 56
6.2 Future Work 57
6.3 Social Impact 58

References **59**

List of Publications and their Proofs **64**

Plagiarism Report **74**

Curriculum Vitae **84**

LIST OF TABLE[s]

2.1 Evolution of detection methodologies across cyberbullying and network intrusion detection 9
2.2 Analysis of existing cyberbullying detection techniques 12
2.3 Summary of recent machine and deep learning approaches for network intrusion detection. 15

3.1 Class distribution of the Facebook Hateful Meme dataset across the three official splits 19
3.2 Principal feature groups in CICIDS2017 20
3.3 Approximate raw label distribution in merged CICIDS2017 before balancing 21
3.4 Label consolidation from CICIDS2017 sub-categories to five unified classes 24

4.1 Side-by-side comparison of the two proposed architectures. 27

5.1 Training hyperparameters for the cyberbullying detection framework . 37
5.2 Training hyperparameters for the network intrusion detection framework 38
5.3 Confusion matrix - Hybrid GAT+BiLSTM+MLP (CICIDS2017) . . . 45
5.4 Per-class AUC scores of Hybrid GAT+BiLSTM+MLP 48
5.5 Performance comparison across models on NIDS 50
5.6 Cross-domain performance summary 53

LIST OF FIGURE[s]

1.1 Cyber threat domains in the digital ecosystem 3

2.1 Comparison of multimodal fusion strategies and the proposed MCB fusion framework 11

2.2 Signature-Based and Anomaly-Based IDS 13

3.1 Parallel dual-branch preprocessing pipeline for cyberbullying detection 21

3.2 Five-stage preprocessing pipeline 23

3.3 Log-scale class distribution before and after random undersampling . 25

4.1 Block diagram of the proposed Cyberbully detection model 27

4.2 DR-HRNet four-stage parallel branch architecture 28

4.3 MC-DeBERTa architecture 31

4.4 Architecture of the proposed hybrid GAT-BiLSTM-MLP intrusion detection framework 33

5.1 Training and Testing Accuracy and Loss Curve 40

5.2 Confusion matrix of Cyberbully Detection 40

5.3 ROC curve of Cyberbully Detection 41

5.4 Accuracy of Cyberbully Detection 42

5.5 Precision of Cyberbully Detection 43

5.6 Recall of Cyberbully Detection 43

5.7 F1 Score of Cyberbully Detection 44

5.8 Confusion Matrix (Proposed model) 47

5.9 ROC Curve (Proposed model) 48

5.10 Performance Metrics accross models on NIDS 49

5.11 Performance evaluation of the GNN model. 50

5.12 Performance evaluation of the CNN model. 51

5.13 Performance evaluation of the MLP model. 51

5.14 Performance evaluation of the LSTM model. 52

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

Symbols	**Description**
α	Attention weight (GAT); Focal Loss weight
α_{ij}	Normalised GAT attention weight from node j to node i
α_c	Class weight for class c, defined as $\alpha_c = N/\hat{n}_c$
γ	Focal loss focusing parameter ($\gamma = 2.0$)
$\delta_k(\cdot)$	k-th stochastic augmentation transform applied to an image
ε	Small numerical constant for optimiser stability (AdamW)
η	Learning rate in gradient-based optimisation
λ	L2 weight-decay coefficient in AdamW
μ_j	Feature-wise mean used for StandardScaler normalisation
σ	Gaussian kernel deviation; sigmoid activation function
σ_j	Per-feature training-set standard deviation in StandardScaler
ρ	ReduceLROnPlateau learning-rate reduction factor ($\rho < 1$)
θ_t	Model parameter vector at optimisation step t
$\oplus$	Feature concatenation operator
$\odot$	Element-wise (Hadamard) multiplication
$*$	Convolution operator
$\|\cdot\|_2$	ℓ_2 (Euclidean) norm
$\top$	Vector/matrix transpose
$\sum$	Summation operator
$\prod$	Product operator
$\arg\max$	Argument of the maximum
$\tanh(\cdot)$	Hyperbolic tangent activation function
$\mathrm{sm}(\cdot)$	Softmax function

Abbreviations	**Full Form**
ACK	Acknowledgement (TCP flag)
AdamW	Adam with Decoupled Weight Decay
APT	Advanced Persistent Threat
AUC	Area Under the (ROC) Curve
BENIGN	Normal (non-attack) network traffic class
BERT	Bidirectional Encoder Representations from Transformers
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network

CICIDS2017	Canadian Institute for Cybersecurity IDS Dataset 2017
CICFlowMeter	Network flow feature extraction tool developed by CIC
CLIP	Contrastive Language–Image Pre-training
CLS	Classification token added to transformer input sequence
CNet	CentralNet multimodal fusion architecture
CSV	Comma-Separated Values
DBN	Deep Belief Network
DeBERTa	Decoding-enhanced BERT with Disentangled Attention
DDoS	Distributed Denial of Service
DINOv2	Self-supervised Vision Transformer (version 2)
DNN	Deep Neural Network
DoS	Denial of Service
DPI	Deep Packet Inspection
DR-HRNet	Dense Residual High-Resolution Network
DRCB	Dense Residual Convolutional Block
DTU	Delhi Technological University
EfficientNet	Efficient Neural Network (scalable CNN architecture)
FHM	Facebook Hateful Meme (dataset)
FIN	Finish flag (TCP control signal)
FN	False Negative
FP	False Positive
GAT	Graph Attention Network
GATConv	Graph Attention Convolution layer
GIN	Graph Isomorphism Network
GNN	Graph Neural Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HRNet	High-Resolution Network
HTTPS	Hypertext Transfer Protocol Secure
IAT	Inter-Arrival Time
IDS	Intrusion Detection System
IoT	Internet of Things
IR	Imbalance Ratio
JSON	JavaScript Object Notation
kNN	k-Nearest Neighbour
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory

MC-DeBERTa	Multi-Scale DeBERTa
MCB	Multimodal Compact Bilinear (Pooling)
MDCB	Multi-Scale Dilated Convolutional Block
MLP	Multi-Layer Perceptron
MobileNetV2	Lightweight convolutional neural network architecture v2
NaN	Not a Number
NB	Naïve Bayes
NIDS	Network Intrusion Detection System
NSL-KDD	NSL Knowledge Discovery and Data Mining dataset
PSH	Push flag (TCP control signal)
RF	Random Forest
ReLU	Rectified Linear Unit
ResNet	Residual Network
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RoBERTa	Robustly Optimised BERT Pre-training Approach
RST	Reset flag (TCP control signal)
SEP	Separator token in transformer input sequences
SHAP	SHapley Additive exPlanations
SSL	Secure Sockets Layer; also Self-Supervised Learning
SVM	Support Vector Machine
SYN	Synchronise flag (TCP control signal)
TCP	Transmission Control Protocol
TLS	Transport Layer Security
TN	True Negative
TP	True Positive
TSP	Transformer Spatial Pooling (module)
UTF-8	Unicode Transformation Format – 8-bit
VGG	Visual Geometry Group (CNN architecture)
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
XLM-RoBERTa	Cross-lingual RoBERTa model
XMTF	Cross-Modal Transformer Fusion framework

Nomenclature	**Meaning**
x_i	Feature vector of the i-th network flow (node)

$\tilde{x}_{ij}$	Normalised value of feature j for sample i
X	Preprocessed feature matrix, $X \in \mathbb{R}^{N\times 73}$
y	Ground-truth label vector, $y \in \{0,1,2,3,4\}^N$
$\hat{y}_t$	Predicted class label for sample t
J	Raw meme image
J_R	Resized image, $(h,w) = (224,224)$
J_N	Denoised image after Gaussian filtering
J_A	Augmented image used for training
G_σ	Gaussian kernel with standard deviation σ
T	Raw meme text caption
W	WordPiece token sequence $\{w_1, w_2, \ldots, w_n\}$
w_j	j-th token in the input sequence
e_j	Combined embedding for token w_j
e_j^t	Token (content) embedding for w_j
e_j^p	Segment type embedding for w_j
e_j^R	Relative position embedding for w_j
C_h	DeBERTa hidden-state output, $C_h \in \mathbb{R}^{n\times d}$
C_{h_D}	Output of MDCB dilated convolution at dilation rate D
D	Dilation rate in Multi-Scale Dilated Convolutional Block
F_{img}	Spatial image feature vector produced by DR-HRNet
F_{txt}	Textual feature vector produced by MC-DeBERTa
f_{fusion}	Joint multimodal feature vector after MCB Pooling
$f_S^{(j)}$	Feature map of the j-th branch at stage S in DR-HRNet
$\tilde{f}_S^{(q)}$	Fused feature map at resolution q after transition module
$T_{j\to q}$	Resolution transformation from branch j to resolution q
Q	Number of parallel branches in DR-HRNet
Y_0	Concatenated input to the DRCB
Y_l	Output of the l-th convolutional layer in DRCB
F^{final}	Final image feature vector output of DRCB
L	Number of convolutional layers in DRCB
$\text{Tr}(\cdot)$	Disentangled Transformer encoder function
W_D, B_D	Dilated convolution weight and bias at rate D in MDCB
f_e	EfficientNetV2 deep feature map, $f_e \in \mathbb{R}^{h\times w\times c}$
$\varepsilon_{\text{effv2}}(\cdot)$	EfficientNetV2 backbone feature extraction function
q,k,v	Query, key, and value matrices in self-attention
$\text{Att}(q,k,v)$	Scaled dot-product attention output in TSP module

o	Logit vector before softmax, $o \in \mathbb{R}^K$
o_k	Logit for class k
K	Number of output classes
$p(Y=k)$	Softmax probability of class k
$\mathscr{G}$	Flow similarity graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$
$\mathscr{V}$	Node set of the flow graph (each node = one network flow)
$\mathscr{E}$	Edge set of the flow graph
k	Number of nearest neighbours in kNN graph construction
$\cos\_\mathrm{sim}(x_i, x_j)$	Cosine similarity between flow feature vectors x_i and x_j
e_{ij}	Raw (unnormalised) attention score for edge $j \to i$ in GAT
a	Learnable attention weight vector in GAT
W	Learnable linear transformation matrix in GAT
h_i	Hidden representation of node i
h'_i	Updated node representation after GAT aggregation
$h_i^{(1)}$	Node embedding after first GATConv layer
$\mathscr{N}(i)$	Neighbourhood set of node i in the flow graph
K_{heads}	Number of attention heads in multi-head GAT ($K = 8$)
H_{GAT}	GAT node embedding matrix, $H_{\text{GAT}} \in \mathbb{R}^{N \times d_{\text{GAT}}}$
f_t	Forget gate activation in LSTM at time step t
i_t	Input gate activation in LSTM at time step t
o_t	Output gate activation in LSTM at time step t
c_t	Cell state of LSTM at time step t
h_t	Hidden state of LSTM at time step t
$\overrightarrow{h}_t$	Forward hidden state of BiLSTM at time step t
$\overleftarrow{h}_t$	Backward hidden state of BiLSTM at time step t
$h_t^{(\text{BiLSTM})}$	Concatenated BiLSTM output at time step t
W_f, U_f, b_f	Forget-gate weight, recurrent weight, and bias in LSTM
W_i, U_i, b_i	Input-gate weight, recurrent weight, and bias in LSTM
W_c, U_c, b_c	Cell-state weight, recurrent weight, and bias in LSTM
W_o, U_o, b_o	Output-gate weight, recurrent weight, and bias in LSTM
$z_t^{(1)}, z_t^{(2)}$	Hidden activations of MLP layers 1 and 2
W_1, b_1	Weight matrix and bias of MLP layer 1
W_2, b_2	Weight matrix and bias of MLP layer 2
$p_{t,c}$	Predicted probability of class c for sample t
$\mathscr{L}_{\text{FL}}$	Focal Loss function
N	Total number of training samples

n_c	Number of samples belonging to class c
$\hat{n}_c$	Number of samples in class c after random undersampling
$N_{\max}$	Maximum per-class sample cap for random undersampling
n_{majority}	Sample count of the majority class
n_{minority}	Sample count of the minority class
IR	Imbalance Ratio $= n_{\text{majority}}/n_{\text{minority}}$
$\hat{m}_t, \hat{v}_t$	Bias-corrected first and second moment estimates in AdamW
C	Total number of traffic classes ($C = 5$ for NIDS)
d	Hidden dimension size of transformer encoder
D	Attention head dimension in self-attention
n	Sequence length (number of tokens) in text branch
h, w, c	Height, width, and channel depth of a feature map

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The accelerated digitisation of technology has revolutionised the way that human beings communicate, conduct business and manage critical infrastructure. Billions of people are interacting everyday via social media, cloud and cloud services, networked enterprise systems, and producing an unprecedented level of digital data . This digital revolution carries with it great socioeconomic advantages, but it has at the same time, laid the groundwork for a variety of cyber attacks [1, 2]. Two categories of such threats have emerged as particularly pressing concerns for researchers, policymakers, and practitioners alike: *cyberbullying* in online social spaces and *network intrusions* targeting digital infrastructure.

Cyberbullying is when someone uses digital technology to mean to cause harm, or does so over and over again. Harass, threaten, demean or harm people (through abusive texts), Hostile memes, offensive images and coordinated online attacks [3, 4]. According to recent world estimates, at least one out of three young people in the world has been exposed to the internet form of bullying that negatively affects academic performance and/or mental health [5]. The anonymous and viral nature of social media increases the psychological effects of this type of abuse, which requires intelligent detection systems that are able to analyse complex multimodal content at scale [6].

Network intrusion, on the other hand, is accessing or attacking a network without permission. Denial of Service (DDoS), brute force, and Advanced Persistent Threats (APTs) activity directed at computer networks [2]. Check Point Research reported a 38% rise in global cyberattacks between 2021 and 2022, with organisations facing an average of more than 1,200 attacks per week. The rise in the use of encryption further makes detection difficult as the traditional methods would not be effective. Deep Packet Inspection (DPI) techniques can't stop encrypted traffic payloads [7, 8].

The similarity in the way cyberbullying comes about with network intrusion in

terms of application structure is that in both cases the adversary embeds malicious intent in legitimate looking data hateful memes which appear funny on the surface, or attack traffic that looks like benign network flows. So, a common deep learning framework to attack the two threats is of important scientific and practical interest.

Our work is motivated by three important insights . First, existing cyberbullying detection systems employ unimodal analysis either text or image [9, 10]. They disregard the cross-modal semantic interactions found in today's multimodal hate speech. Secondly, conventional Network Intrusion Detection Systems (NIDS) assume the network flows as independent and identically distributed objects. They overlook the rich topological and temporal structure underlying coordinated attacks [11, 12].

This work bridges that gap by proposing and evaluating two complementary hybrid deep learning systems: a multimodal cyberbullying detector built upon Dense Residual High-Resolution Networks (DR-HRNet) and Multi-Scale DeBERTa (MC-DeBERTa) with Multimodal Compact Bilinear (MCB) pooling and Adaptive EfficientNetV2 and a graph-based network intrusion detector that combines GAT, BiLSTM, and MLP classifiers. Together, these systems demonstrate that hybrid deep learning architectures can achieve robust, high-accuracy cyber threat detection across fundamentally different threat domains.

1.2 The Rise of Cyber Threats in the Modern Times

Cyber threats have evolved in such a way as to closely match the evolution of digital technology. Cloud ecosystems, as well as devices connected via Internet-of-Things (IoT), are also increasingly used for communication. The number of attacks against and the sophistication of such attacks has increased, as have their variety. Figure 1.1 shows digital ecosystem.

The attacks are becoming increasingly sophisticated and diverse as we see the rise of communication platforms, cloud ecosystems, and internet-of-things devices.

1.2.1 Cyberbullying as a Social Cyber Threat

Cyberbullying is a type of social aggression. It takes advantage of the anonymity and reach of digital platforms. Online harassment is different from the usual bullying because the information shares quickly and the effects are long-lasting [4] . Single modality detection systems have been challenged by hateful memes. An image and a caption can be two innocuous things, but they can also be hateful when put together [13,

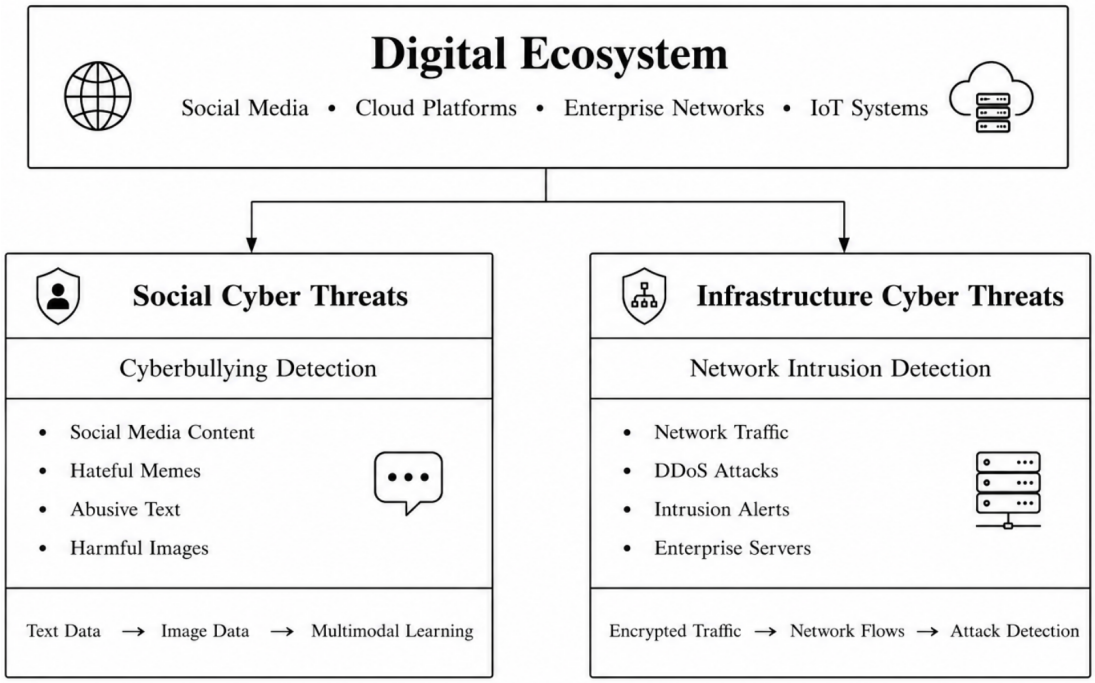


Figure 1.1: Cyber threat domains in the digital ecosystem

14].

Informal and changing online language makes it difficult to detect. There's more to it than sarcasm, coded references and unbalanced datasets [15]. Initial approaches were using SVM, NB, and RF. These features were handcrafted but didn't capture the true meaning of the context [16].

BERT and DeBERTa boosted the ability to detect text. The ResNet and EfficientNet models improved the visual classification. However, in unimodal systems there is no interaction between text and images. The gap sparked increasing interest in the area of detection of multimodal cyberbullying [10, 17].

1.2.2 Network Intrusion as an Infrastructure Cyber Threat

One of the greatest threats and most economically damaging forms of network intrusion is denial of service.The DoS is the most common and costly network intrusion. Intrusions range from a wide variety of attack methods DDoS flooding, port scanning, DoS attack, brute-force. Each has a different network traffic signature and each has different levels of difficulty for automated detection [18, 19]. Traditional IDSs used to do signature-based matching. But it isn't enough to be able to defend against zero-day exploits and polymorphic attacks, which are constantly changing [20].

The increasing use of transport-layer encryption (TLS/SSL) introduces a problem with it. Additional challenge: Encrypted traffic is now a larger percentage of the total traffic. The analysis of network flows is no longer possible for the most part of network traffic at the payload level for majority of traffic. This type of Anomaly-based NIDS works on the statistical features of the flow, such as the number of packets, inter-arrival times, etc., have thus come into the limelight [7, 21]. Machine Learning models and Deep Learning models such as, LSTM networks, convolutional neural networks and autoencoders have been shown to be effective to these flow-level features with considerable success [11, 22, 23].

A major issue in almost all the NIDS benchmark datasets is the severe class imbalance. This imbalance causes standard classifiers to exhibit a strong bias towards the majority class [24]. In order for NIDS to be effective, it must include both The architecture-level innovations and data-level strategies to address the imbalanced network traffic.

1.3 Problem Statement

Given the increasing sophistication and multimodal nature of cyber threats across both social media and network infrastructure domains, existing detection approaches suffer from three fundamental deficiencies:

Problem 1: Unimodal representation. Current cyberbullying detectors fail to capture cross-modal interactions between image and text in hateful content, while NIDS models fail to exploit the graph-topological and temporal structure of network flows.

Problem 2: Inadequate feature fusion. Existing multimodal fusion strategies rely on simple concatenation or shallow attention that cannot model higher-order inter-modal dependencies.

Problem 3: Poor performance under class imbalance. Standard classifiers exhibit high false-negative rates for minority-class threats (hateful content and rare attack types) due to inadequate handling of severe class imbalance in real-world datasets.

1.4 Research Objectives

RO1. Multimodal Cyberbullying Detection: To design and implement a hybrid multimodal deep learning architecture that simultaneously extracts rich spatial features from images using a Dense Residual High-Resolution Network (DR-HRNet)

and deep semantic-contextual features from text using a Multi-Scale DeBERTa (MC-DeBERTa) model, and fuses them using Multimodal Compact Bilinear (MCB) pooling for accurate binary classification of cyberbullying content.

RO2. Graph-Based Network Intrusion Detection: To construct a hybrid graph-based deep learning framework that models network flows as nodes in a similarity graph, learns spatial interaction patterns using multi-head Graph Attention Network (GAT) layers, captures temporal behavioural dependencies using a Bidirectional LSTM (BiLSTM), and classifies traffic into five categories using a Multi-Layer Perceptron (MLP) with class-weighted loss.

RO3. Handling Class Imbalance: To evaluate and implement effective strategies for mitigating class imbalance in both domains, including data augmentation, controlled undersampling, and algorithm-level weighted loss functions, with the goal of improving minority-class recall.

RO4. Empirical Validation: To evaluate both proposed systems on publicly available benchmark datasets the Facebook Hateful Meme Dataset and the CICIDS2017 network intrusion dataset using standard performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC, and to compare results against relevant state-of-the-art baselines.

RO5. Cross-Domain Analysis: To derive insights from both experimental systems regarding the advantages of hybrid deep learning architectures, the role of attention mechanisms, and the shared challenges of generalisation and class imbalance across cyber threat detection domains.

1.5 Key Contributions of the Work

Contribution 1: Novel Multimodal Cyberbullying Detection Framework. A hybrid end-to-end architecture combining DR-HRNet for multi-scale spatial image feature extraction, MC-DeBERTa for multi-scale contextual text feature extraction, MCB Pooling for higher-order cross-modal feature fusion, and Adaptive EfficientNetV2 with a Transformer Spatial Pooling (TSP) module for classification. The proposed framework achieves **99.20% - 99.50%** accuracy and an F1-score exceeding 99% on the Facebook Hateful Meme Dataset, outperforming state-of-the-art baselines.

Contribution 2: Novel Graph-Based Hybrid NIDS Framework. A hybrid architecture that represents network flows as nodes in a cosine-similarity kNN graph,

learns topological attack patterns using two stacked multi-head GATConv layers with batch normalisation and residual connections, models temporal flow evolution using a BiLSTM, and classifies flows using an MLP with class-frequency-weighted Focal loss. The proposed system achieves **94.2%** accuracy and F1-score with a macro ROC-AUC of **0.997** on the CICIDS2017 dataset, surpassing all evaluated baselines (GNN, CNN, MLP, LSTM).

Contribution 3: Systematic Handling of Class Imbalance. A comprehensive empirical study of class imbalance mitigation strategies in both domains, demonstrating that the combination of controlled random undersampling at the data level and class-frequency-weighted loss at the algorithm level yields superior minority-class detection performance.

Contribution 4: Multi-Domain Comparative Analysis. A unified cross-domain analysis that identifies shared principles governing the effectiveness of hybrid deep learning architectures across structurally distinct cyber threat domains, providing insights applicable to future multi-domain threat detection research.

1.6 Thesis Organisation

The remaining sections of this work are structured as follows:

Chapter 2 provides a detailed review of previous work in both cyberbullying detection and network intrusion detection, covering classical machine learning methods, CNN/RNN-based deep learning approaches, transformer-based models, multimodal fusion systems, and graph-neural-network-based methods. Research gaps motivating the proposed contributions are identified at the end of the chapter.

Chapter 3 describes the two benchmark datasets used in this work the Facebook Hateful Meme Dataset and the CICIDS2017 network intrusion dataset along with the preprocessing pipelines applied to each. Dataset statistics, class distributions, and the rationale for design choices in preprocessing are presented.

Chapter 4 presents the technical details of both proposed systems. For Domain 1, the DR-HRNet image feature extractor, MC-DeBERTa text feature extractor, MCB pooling fusion mechanism, and Adaptive EfficientNetV2 classifier are described with supporting mathematical formulations. For Domain 2, the kNN graph construction procedure, multi-head GAT layers, BiLSTM temporal modelling, MLP classification head, and training strategy are detailed.

Chapter 5 is a complete empirical evaluation of both systems presented. The analysis includes a confusion matrix, ROC curves and performance metrics such as accuracy and F1 score, precision and recall along with the training convergence curves of all models which are compared with the baselines. The results of both domains are described next. This section deals with why hybrid architectures better than single paradigm alternatives, role of attention mechanism, how class imbalance was handled and at what cost, computational complexity and overhead of each framework, and finally shortcomings of the proposed work.

Chapter 6: concludes the work by giving the summary of contributions, findings, social impact and ethical issues, limitations and scope for future research.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews existing literature across two domains central to this work: cyberbullying detection and network intrusion detection. Rather than cataloguing every published work, the goal is to trace how methods have evolved in each domain, identify what has and has not been solved, and explain where the proposed frameworks fit in.

This chapter reviews the current research work in two domains that are very important to the work presented in this thesis. This is cyberbullying detection and network intrusion detection. We do not try to enumerate all the published work but trace what methods were tried, what was solved and what remained unsolved, and where our proposed frameworks sit.

2.1 Overview

Cyber threat detection includes several security challenges such as flagging cyber-bullying posts and identifying suspicious network traffic in enterprise networks.

What connects all of them. distinguishing adversarial or harmful behaviour from legitimate is the chief difficulty. Manoeuvre in loud, high-volume information data streams A broad range of problems are covered by Cyber Threat Detection from flagging malicious social. Finding malicious packets in enterprise networks [6, 18].

Threats can be categorised into two types at a high level. When it comes to Research work in law software and engineering under this topic which was recently done. Network-based threats are attacks that are powered through resources in the communication infrastructure. Examples include denial of service attack, port scan, intrusion attempts, and malware.

Up to now, the preferred methods in both domains have been somewhat different, but highly effective hybrid architectures that use building blocks from various learning methods [1, 25, 26]. Table 2.1 highlights the key approaches for the detection in text

and image domain till date.

Table 2.1: Evolution of detection methodologies across cyberbullying and network intrusion detection

Era	**Cyberbullying**	**NIDS**	***Key Limitation***
Rule-based (∼2008–2013)	Keyword lists, blacklists	Snort, signature matching	*Cannot detect unknown threats*
Classical ML (∼2013–2017)	SVM, Naïve Bayes, Random Forest	SVM, Decision Tree, kNN	*Manual feature engineering*
Shallow DL (∼2017–2020)	CNN, LSTM, GRU	LSTM, CNN, Autoencoder	*Single modality; limited context*
Transformers (∼2020–2022)	BERT, RoBERTa, DeBERTa, ViT	Attention-based anomaly detection	*High compute; unimodal*
Multimodal / Graph (∼2022–2024)	CLIP, CentralNet, MCB Fusion	GNN, GAT, Graph autoencoders	*Shallow fusion; imbalance issues*
Hybrid (current)	**DR-HRNet + MC-DeBERTa + EfficientNetV2**	**GAT + BiLSTM + MLP**	*This work*

2.2 Cyberbullying Detection

Research in cyberbullying detection has broadly divided into five phases. We trace that trajectory here, stressing what was achieved and what was not.

2.2.1 Classical Machine Learning Approaches

The early work introduced hand-crafted features such as word n-grams and bag-of-word vectors that were fed into a classifier. Support vector machine. Naïve Bayes algorithm. Random forest [6]. SVM was widely used as it is effective in high dimensional text spaces but were reported to degrade rapidly on the more informal and misspelling heavy language found on social media [15]. While using these methods, the more fundamental difficulty involved structure since they required surface-level lexical signals, which were more likely to miss things like irony, sarcasm, coded language, and the visual [16].

2.2.2 CNN and RNN Based Methods

Deep learning eliminates manual feature engineering requirements. Applying a convolutional neural network (CNN) to text enables the use of 1-D convolutions to capture local n-gram patterns. In these images, you will find CNNs originally designed for images (VGG, ResNet, etc.) [27]. Arshed *et al.* [28] introduced the CVID visual cyberbullying dataset and showed that DenseNet201 reached 99% accuracy on it. On the text side, Tabassum and Nunavath [29] compared nine models across image and text branches, finding that ViT reached 99.5% on image data and RoBERTa reached 99.2% on text but both remain unimodal, so they simply cannot detect cross-modal hateful content that characterises real-world memes. Hybrid CNN-RNN architectures that combined local and sequential modelling showed incremental gains but ran into the same single-modality ceiling [30].

2.2.3 Transformer-Based Methods

BERT and its variants changed text-based cyberbullying detection considerably. Bidirectional pretraining on large corpora gave these models the contextual understanding needed to handle figurative language and implicit threats [31]. RoBERTa improved on BERT through better pretraining strategies [29], and DeBERTa went further by separately encoding content and positional information through disentangled attention particularly useful for the short, syntactically irregular sentences that dominate social media [9, 32]. On the visual side, Vision Transformers applied patch-based self-attention to image classification. Kumari and Kaur [3] proposed XMTF, combining DINOv2 visual features with a transformer text encoder for code-mixed meme analysis. The persistent limitation is that most of these systems are unimodal, and fusing two transformer-based representations across modalities remains an open problem [14, 33, 34].

2.2.4 Multimodal Fusion Approaches

The community has largely converged on three fusion strategies: *early* (concatenating raw features before encoding), *late* (encoding each modality independently and combining decisions), and *intermediate* (interacting between modality-specific representations mid-encoding). Figure 2.1 illustrates these strategies, including the MCB-based approach proposed in this work.

Mohammedjany *et al.* [9] combined VGG16 with XLM-RoBERTa through attention, CLIP alignment, and CentralNet fusion, but only managed 74% accuracy a

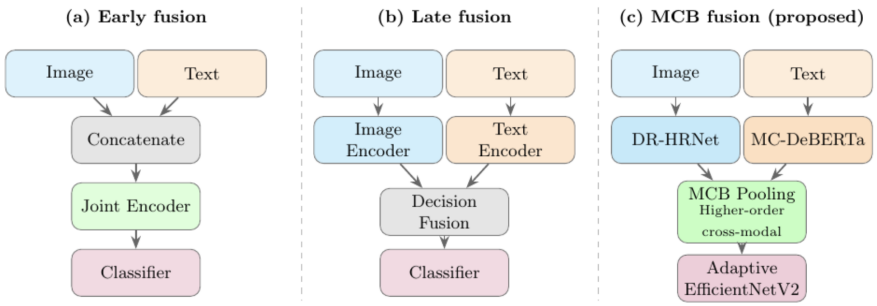


Figure 2.1: Comparison of multimodal fusion strategies and the proposed MCB fusion framework

clear sign that the fusion strategy itself was the bottleneck. Singh *et al.* [10] used a BiLSTM-AHCNet for text, a tuned EfficientB0 for images, and a horse herd optimisation weighting scheme to combine predictions, but simple weighted late fusion still leaves cross-modal interactions unaddressed. Wandeto *et al.* [31] concatenated BERT and ResNet50V2 features directly. Luo *et al.* [17] took a different approach entirely, proposing RSBully a reinforced multi-agent subgraph neural network that exploits the social network graph structure rather than post-level image-text content, achieving strong results but requiring access to the full relational graph of users and posts.

2.2.5 Recent Works (2024–2025)

The past two years have seen a wider spread of approaches. Phosit *et al.* [33] tackled Thai-language memes, demonstrating the need for language-specific adaptations in the text encoder. Yang *et al.* [14] proposed CyberGenDet for cross-lingual detection and content generation across multiple language pairs. Jha *et al.* [13] added multimodal explainability on top of detection, identifying which image regions and text phrases drive a bullying prediction an important step for content moderation at scale. Li *et al.* [35] introduced GIN-based multimodal feature transformation with multi-feature voting targeting irony-aware detection. Tahir and Nawaz [4] proposed HUEID for Urdu text incorporating emoji signals, and Mahmud and Roy [5, 36] extended multimodal detection to Romanized Bangla. Despite this range of work, most systems still rely on simple concatenation or late fusion, few explicitly model higher-order cross-modal interactions, and class imbalance handling remains inconsistently reported across the literature. Table 2.2 summarises the key cyberbullying works reviewed above.

Table 2.2: Analysis of existing cyberbullying detection techniques

Author & Year	**Dataset**	**Techniques**	**Limitation**
Tabassum *et al.* [29] (2024)	Self-collected & Public	ViT, RoBERTa, CNN, LSTM	Unimodal; no cross-modal fusion
Mohammedjany *et al.* [9] (2025)	Social media meme dataset	VGG16, XLM-RoBERTa, CLIP	Shallow fusion; low accuracy
Singh *et al.* [10] (2024)	Multimodal social media	BiLSTM-AHCNet, EfficientB0	Simple weighted fusion
Arshed *et al.* [28] (2025)	CVID dataset	DenseNet201, VGG, ViT	Visual only; no text modality
Luo *et al.* [17] (2025)	Real-world social media	RSBully (Graph NN)	Graph-structure dependent
Wandeto *et al.* [31] (2024)	Custom multimodal	BERT + ResNet50V2	Naive feature concatenation
Kumari & Kaur [3] (2025)	Code-mixed memes	XMTF (DINOv2 + Transformer)	Language specific

2.3 Network Intrusion Detection Systems (NIDS)

The research on NIDS is posterior to that on cyberbullying detection. It has been suffering from the same problem for a long time; how to create models that are reliable enough to catch new attack patterns but not so unreliable that they generate so many false alarms that the operators cannot act upon them adequately. In this section we shall look at the .

2.3.1 Signature-Based vs Anomaly-Based IDS

Signature-based IDS tools like Snort and Suricata match traffic against a known-attack database. They're reliable for threats people have seen before [2]. The problem is structural: a zero-day leaves no signature to match, and every time attackers change their approach, someone has to update the rules manually [18].

Anomaly-based IDS take the opposite approach: build a model of normal traffic and flag deviations from it. This gives them some ability to catch novel attacks, but introduces a practical difficulty any legitimate traffic that falls outside the training distribution triggers an alert, leading to high false-positive rates [20]. Figure 2.2

illustrates the two architectures.

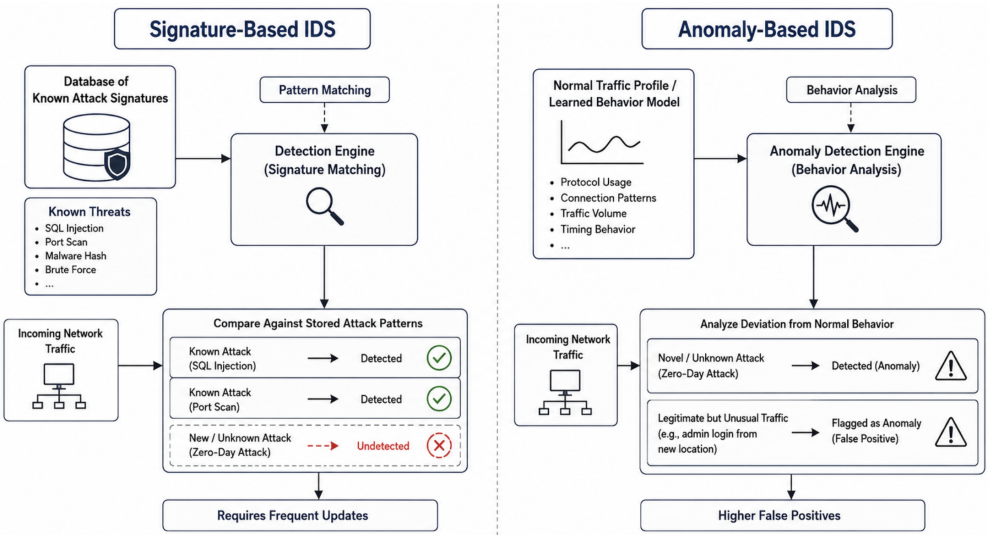


Figure 2.2: Signature-Based and Anomaly-Based IDS

2.3.2 Machine Learning and Deep Learning Approaches

Supervised ML brought systematic performance gains to anomaly-based detection. ALzaher and AlJarullah [20] sconducted an overview of intrusion detection using machine learning in supervised, unsupervised, and hybrid approaches. They reflected upon previous studies by mentioning how effective popular supervised models like SVM and Decision Trees are within the domain of known attacks. At the same time, unsupervised methods allow zero-day detection through behavioral analysis. Tree-based models achieved strong classification accu- racy and SHAP improved interpretability by identifying influential traffic features [37].

For unsupervised detection, Mozaffari *et al.* [38] showcased an anomaly detection system accessible on the web that can dynamically adjust to real-time high-dimensional network streams through the utilization of supervised, semi-supervised, and self-supervised learning. Sattar *et al.* [7] presented ET-SSL, an autonomous contrastive learning algorithm for detecting anomalies in concealed settings. This model uses packet length, inter-arrival time, and flow duration to detect abnormalities successfully across datasets such as CIC-Darknet2020 and UNSW-NB15. It is scalable and robust in detecting zero-day assaults. Ferrag *et al.* [22] benchmarked a range of architectures RNN, DNN, CNN, DBN, autoencoders across 35 datasets, establishing CNNs as a reliable baseline. Yin *et al.* [11] presented LSTM networks for the detection of stealthy and evolving attacks. Sajid *et al.* [39] hybrid models improved sequential learning and achieved higher detection rates than standalone ML/DL models. Alsoufi *et al.* [40]

extended autoencoder-based detection to IoT environments. Kim and Lee [12] created a deep autoencoder and classifier hybrid architecture that outperformed all others on benchmarked datasets. Regardless, there are ongoing problems with high false-positive rates, restricted real-time deployment capabilities, and low generalizability to imbalanced datasets.

2.3.3 Graph Neural Network Approaches

Fu *et al.* [41] graph-based modeling captured traffic interactions and achieved strong detection results without requiring labeled data. Ferriyan *et al.* [42] built the HIKARI-2021 encrypted traffic dataset and ran GNN-based methods on it. Graph representations found structural attack signatures that flat feature vectors simply don't see. Ibraheem *et al.* [21] showed that careful feature selection improves ML-based anomaly detection in HTTPS traffic, though without exploiting graph structure.

Graph Attention Networks extend the basic GNN by learning to weight each neighbour's contribution based on feature compatibility. This is well-suited to NIDS, where the relevance of neighbouring flows varies with the attack type. The proposed framework uses two stacked multi-head GATConv layers with batch normalisation and residual connections, combined with BiLSTM to capture temporal flow sequences simultaneously.

2.3.4 Encrypted Traffic, Flow-Level Detection, and Class Imbalance

The shift to HTTPS and TLS 1.3 has made payload inspection increasingly futile for a growing share of real-world traffic. Papadogiannaki and Ioannidis [8] survey the move to statistical flow-level analysis as the dominant detection approach for encrypted traffic. Elmaghraby *et al.* [43] showed that flow-level features alone support competitive detection accuracy on encrypted benchmark datasets, validating the approach taken in this work. The CICIDS2017 dataset [44] is built around this paradigm, using CICFlowMeter-extracted flow statistics without any payload content.

Class imbalance is the most persistent practical obstacle in NIDS evaluation. Benign traffic routinely accounts for 80% or more of any real-world dataset, pushing standard loss functions to favour the majority class and producing high false-negative rates for rare attack categories [24]. Goldschmidt and Chudá [24] analysed 89 IDS datasets and identified imbalance as the single most damaging data quality issue, recommending combined data-level and algorithm-level mitigation. This work uses random undersampling at the data level together with class-frequency-weighted cross-

Table 2.3: Summary of recent machine and deep learning approaches for network intrusion detection.

Author & Year	Dataset	Models	Limitation
Ferrag *et al.* [22] (2020)	35 datasets; CICIDS2018	RNN, DNN, CNN, Autoencoder	Limited generalisation
Singh *et al.* [37] (2025)	CICIDS, USTC-TFC, CSE	XGBoost + SHAP	No temporal/graph modelling
Fu *et al.* [41] (2023)	92 real-world datasets	Graph-based GNN	Computationally expensive
Ferriyan *et al.* [42] (2021)	HIKARI-2021	Encrypted dataset + DL	Synthetic attacks only
Sajid *et al.* [39] (2024)	CICIDS, UNSW, NSL-KDD	XGBoost+LSTM, CNN+LSTM	No graph; flow-independent
Sattar *et al.* [7] (2025)	CIC-Darknet2020; UNSW-NB15	ET-SSL (contrastive)	Unsupervised; no class labels
Yin *et al.* [11] (2017)	KDD Cup 99; NSL-KDD	LSTM (RNN)	Unimodal sequential; no graph

entropy loss at the algorithm level a dual strategy that ensures balanced gradient updates across all five traffic categories.

Table 2.3 summarises the NIDS works reviewed above.

2.4 Hybrid Deep Learning in Cybersecurity

No single learning paradigm handles every aspect of modern cyber threats, which has pushed researchers toward frameworks that combine convolutional, recurrent, attention, and graph modules in a single end-to-end system. In NIDS, hybrid CNN+LSTM models were an early step; Sajid *et al.* [39] showed they outperform either component alone, but without graph structure they still treat flows independently. Adding GNN layer to capture inter-flow topology. Which combined with BiLSTM to model sequential flow behaviour, provides coordinated attacks . This is the core idea behind the GAT-BiLSTM-MLP framework proposed here, which achieved 94.2% F1 on CICIDS2017 compared to 90.65% for a standalone GNN and 59.53% for a standalone LSTM.

In cyberbullying detection, the hybrid philosophy involves pairing specialist

image encoders with specialist text encoders through a fusion mechanism capable of modelling cross-modal interactions, rather than simply concatenating outputs. The DR-HRNet + MC-DeBERTa + MCB + Adaptive EfficientNetV2 framework realises this: multi-resolution spatial features from the image branch and multi-scale contextual features from the text branch are fused through MCB pooling, which approximates the full bilinear outer product between modalities. The system reached 99.20%–99.50% accuracy on the Facebook Hateful Meme Dataset.

2.5 Identified Research Gaps

The review above reveals a consistent set of shortcomings that the proposed frameworks directly address.

Cyberbullying detection gaps. Standard image encoders progressively downsample spatial resolution, discarding fine-grained visual cues that matter for meme-based bullying detection. DR-HRNet preserves high-resolution feature maps throughout all four stages. Most transformer text encoders operate at a single receptive field scale, missing multi-granularity linguistic patterns; MC-DeBERTa adds MDCB dilated convolutions to address this [28]. The dominant fusion strategy across the literature is naive concatenation or simple late fusion; MCB pooling provides higher-order cross-modal interaction modelling that has not previously been applied in this domain. Finally, class imbalance is inconsistently reported many systems quote overall accuracy without a per-class breakdown, which hides poor recall on the minority (hateful) class.

NIDS gaps. Most deep learning NIDS models treat each flow as an independent vector, completely ignoring the inter-flow topology that coordinated attacks produce [41]. Spatial (graph) and temporal (sequence) modelling are rarely combined in a single end-to-end architecture, despite the two being complementary rather than competing. Class imbalance is widely acknowledged but inconsistently handled; the dual strategy of undersampling plus weighted loss used here is underutilised in the literature. Explainability remains an open challenge [45, 46].

Cross-domain observations. This work brings together two independently developed works each addressing a distinct cyber threat domain with its own dataset, architecture, and evaluation setup. The cyberbullying detection work (Chapter 4) and the NIDS work operate on different data types, use different hybrid architectures, and are not designed as a unified joint model. The work combines them to examine whether similar hybrid design principles richer feature representations, stronger fusion, and

explicit handling of class imbalance produce consistent gains across domains that are typically studied in isolation. Real-time deployment and cross-dataset generalisation remain open challenges for both.

The gaps identified here directly motivate the architectural choices described in Chapter 4, and the datasets used to evaluate them are introduced in Chapter 3.

CHAPTER 3

DATASETS AND PREPROCESSING

3.1 Introduction

This encompasses number of cyber threat areas which vary significantly in terms of data structure. Cyberbullying in Social Media memes and Intrusion detection in enterprise network traffic. Every domain requires its own set of data and its own pre-processing pipeline.

For **Cyberbullying Detection**, the Facebook Hateful Meme Dataset [47] is used. It pairs meme images with their overlaid text, making it the standard benchmark for systems that must jointly reason over visual and linguistic content to identify bullying intent [9, 10].

For Cyberbullying Detection, FHM Dataset [47] is used. It is the common benchmark for this feature. Systems that are required to collectively infer meaning from the visual and linguistic information to detect bullying. intent [9, 10].

For **Network Intrusion Detection**, CICIDS2017 dataset [44] is used. Produced at the Canadian Institute for Cybersecurity, using a realistic network setup. It offers flow level statistical capabilities for a wide variety of contemporary attacks and is one of the most well-known benchmarks for intrusion detection [22, 39].

3.2 Cyberbullying Detection Dataset

3.2.1 Facebook Hateful Meme Dataset

Dataset published by Facebook AI Research known as FHM (Hateful Meme) [47]. It is part of an open challenge on multimodal hateful content detection. Each sample is a meme image with the text caption overlaid on it, exactly as it appears on the image. Is found on sites such as Facebook, Twitter and Instagram.

One of its notable characteristics is that it has many samples labeled as benign, a property it displays because its input is confounded [9, 10]. Dataset genuine stress-test for multimodal architectures. It comprises **10,000 samples**, each labelled *Bully* or *Non-Bully*, with images as standardised JPEG files and captions as UTF-8 JSON entries. The dataset was obtained from its Kaggle mirror under non-commercial research terms [29, 47].

3.2.2 Dataset Statistics and Split

Table 3.1 shows the class distribution across the three official splits. Overall, the ratio is approximately **37.6% Bully** to **62.4% Non-Bully**, with the imbalance concentrated in the training set. The development and test sets are near-balanced to allow fair evaluation [29, 47].

Table 3.1: Class distribution of the Facebook Hateful Meme dataset across the three official splits

Split	Bully	Non-Bully	Total	Bully (%)
Training Set	3,019	5,481	8,500	35.5
Development Set	247	253	500	49.4
Test Set	490	510	1,000	49.0
Overall	**3,756**	**6,244**	**10,000**	**37.6**

The three partitions ship pre-defined and non-overlapping [47]. The training set (8,500 samples) utilised for parameter optimisation, the validation set (500 samples) for hyperparameter selection and the test set (1,000 samples) for performance reporting. This fixed split is kept without modification to allow direct comparison with prior work [9, 10, 29].

3.3 Network Intrusion Detection Dataset

3.3.1 CICIDS2017 Dataset

The CICIDS2017 dataset was developed to address the well-known weaknesses of older IDS benchmarks like KDD-Cup 99 and NSL-KDD, which had outdated attack types and unrealistic traffic profiles [18, 44]. It was captured over five consecutive days on a purpose-built testbed, combining realistic background traffic with seven families of modern attack traffic [22, 44].

Covered attack categories include **Brute Force**, **DoS/DDoS**, **Infiltration**, **Botnet**,

Table 3.2: Principal feature groups in CICIDS2017

Feature Group	Representative Features	Description
Flow Duration	Flow Duration (s)	Total elapsed time of the flow
Byte Statistics	Total Length Fwd/Bwd Pkts	Byte-level traffic volume
TCP Flag Counts	FIN, SYN, RST, PSH, ACK flags	TCP control-signal distribution
Packet Length	Pkt Len Mean/Std/Variance	Size heterogeneity of packets
Rate-Based	Flow Pkts/s, Flow Bytes/s	Throughput metrics
Packet Counts	Total Fwd/Bwd Packets	Volume in each traffic direction
Header Lengths	Fwd/Bwd Header Length	Protocol overhead indicators
Identifier Columns	Src IP, Dst IP, Flow ID, Src Port, Timestamp	Removed before training (Sec. 3.5.2)
Inter-Arrival Time	Flow IAT Mean/Std/Min/Max	Timing between consecutive packets
Label	BENIGN, DDoS, PortScan, DoS, ...	Ground-truth traffic category

Heartbleed, **PortScan** and **Web Attacks**,alongside normal (BENIGN) flows. Flow-level features were extracted using CICFlowMeter, yielding **78 numerical attributes** per flow covering packet statistics, inter-arrival times, flag counts, and protocol metadata [39,44]. All daily CSV files were merged into single DataFrame before preprocessing.

3.3.2 Feature Groups and Label Distribution

Table 3.2 lists the principal feature groups. The `Label` column stores the fine-grained traffic category string.

Table 3.3 shows the raw label distribution before preprocessing. The heavy dominance of BENIGN traffic directly motivates the undersampling strategy in Section 3.5.5.

The Imbalance Ratio (IR) between BENIGN and the rarest attack class exceeds 200,000:1 in the raw corpus:

$$\mathrm{IR} = \frac{n_{\text{majority}}}{n_{\text{minority}}} \tag{3.1}$$

This places the dataset firmly in the extreme imbalance regime [22]. The prepro-

Table 3.3: Approximate raw label distribution in merged CICIDS2017 before balancing

Original Label	Approx. Count	Share (%)
BENIGN	2,096,484	83.12
DoS	193,748	7.68
DDoS	128,016	5.08
PortScan	90,819	3.60
Other	13,295	0.52
Total	**2,522,362**	**100.00**

cessing pipeline in Section 3.5 reduces the effective IR to **1**:**1**.

3.4 Preprocessing Pipeline for Cyberbullying Detection

The cyberbullying pipeline runs two independent branches in parallel one for images, one for text before passing their outputs to the respective feature-extraction modules. Figure 3.1 shows the structure.

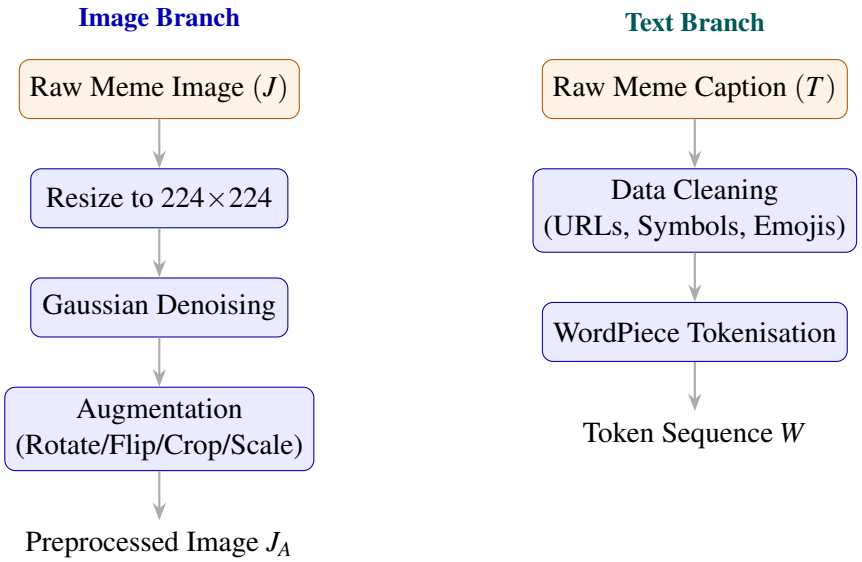


Figure 3.1: Parallel dual-branch preprocessing pipeline for cyberbullying detection

3.4.1 Image Preprocessing

User-generated memes vary widely in resolution, aspect ratio, compression quality, and illumination [47]. Three sequential steps homogenise the input [9, 30].

Resizing. Every image J is rescaled to $(h, w) = (224, 224)$ pixels using bilinear

interpolation, satisfying the DR-HRNet backbone's input requirement:

$$J_R = \text{Resize}(J,\, h,\, w) \tag{3.2}$$

Denoising. A Gaussian low-pass filter suppresses JPEG compression artefacts and other high-frequency noise while preserving structural cues:

$$J_N = G_\sigma * J_R \tag{3.3}$$

where G_σ is Gaussian kernel having standard deviation σ and $*$ denotes convolution.

Augmentation. Stochastic geometric and photometric transformations reduce overfitting and expand the effective training set [9]:

$$J_A = \left\{ \delta_1(J_N),\, \delta_2(J_N),\, \ldots,\, \delta_k(J_N) \right\} \tag{3.4}$$

Applied transforms include random horizontal flipping, rotation by up to $\pm 15^\circ$, uniform scaling, and random cropping.

3.4.2 Text Preprocessing

Social-media captions typically contain informal grammar, emojis, non-standard abbreviations, and URLs [6, 27]. Two steps prepare raw captions for MC-DeBERTa.

Cleaning. URLs and hyperlinks are removed via regex matching. Mention sigils (`@`) and hashtag markers (`#`) are stripped while keeping the embedded tokens. Emojis and special characters are removed, and redundant whitespace is collapsed to single spaces.

Tokenisation. The cleaned caption T is segmented using the WordPiece subword tokeniser with the DeBERTa-v3 vocabulary [9]:

$$W = \{w_1,\, w_2,\, \ldots,\, w_n\} \tag{3.5}$$

Boundary tokens `[CLS]` and `[SEP]` are added, and sequences are padded or truncated to a maximum of 128 tokens before entering MC-DeBERTa's embedding layer.

3.5 Preprocessing Pipeline for Network Intrusion Detection

Preparing raw CICIDS2017 flows for training involves five sequential stages: column standardisation, identifier removal, label grouping, feature normalisation, and class balancing. Figure 3.2 shows the full workflow.

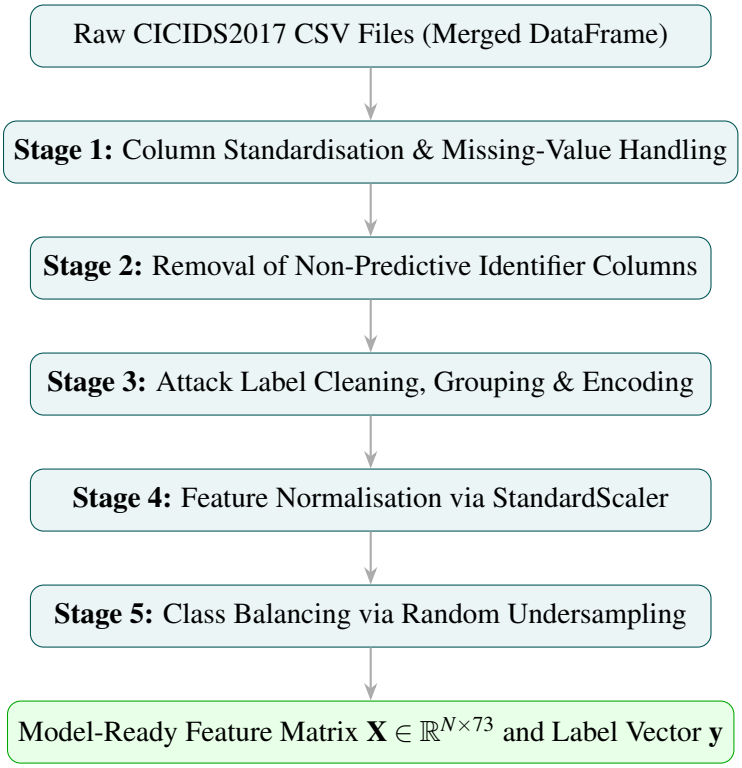


Figure 3.2: Five-stage preprocessing pipeline

3.5.1 Column Standardisation and Missing Value Handling

CICFlowMeter CSV files contain column names with leading and trailing whitespace, which prevents reliable indexing. All column names are stripped before downstream processing [44]. Four data-quality issues are then addressed: infinite values (`+Inf`, `-Inf`) are replaced with `NaN` to prevent gradient explosions; missing numerical values are imputed with zero to preserve row count; rows with a null `Label` field are discarded; and exact duplicate rows are removed to prevent the model from memorising repeated flow fingerprints.

Then four data-quality problems are discussed: infinite values, missing values, outliers and data entry errors. The NaN value is used for missing numerical values to avoid exploding gradients. Zero is inserted when the value is missing. Exact duplicate rows are removed so that model doesn't have to memorize repeated flow fingerprints.

3.5.2 Removal of Non-Predictive Identifier Columns

The following columns are excluded from feature extraction [39, 44] (Flow ID, Source IP, Destination IP, Source Port, and Destination Port). These are specific to the testbed. Eliminating them cuts down the number of features in model from 78 to 73, and if they removed, number of features in model decreases from 78 to 73, Limits learning of model to traffic-behavioural patterns only.

3.5.3 Attack Label Cleaning, Grouping and Encoding

The raw Label field has 14 inconsistent, fine-grained sub-type strings of attacks. whitespace and capitalisation. All labels are normalised and then combined into five. The super-categories as illustrated in Table 3.4. This makes the statistical unlearability of rare a problem that can be resolved. Identifies sub-types and yields a taxonomy that is closer to the taxonomy IDS tools report. threats [18, 22].

Table 3.4: Label consolidation from CICIDS2017 sub-categories to five unified classes

Class	Original Sub-Labels
BENIGN	BENIGN
DDoS	DDoS
PortScan	PortScan
DoS	DoS Hulk, DoS GoldenEye, DoS Slowhttptest, DoS Slowloris
Other	Bot, Web Attack - Brute Force, Web Attack - Sql Injection, Infiltration

3.5.4 Feature Normalisation

All 73 numerical features are normalised using `StandardScaler` [25, 39]:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{3.6}$$

where μ_j and σ_j are the per-feature training-set mean and standard deviation. Normalisation matters here for two reasons. First, gradient-based optimisers converge faster when features share comparable numerical ranges [25]. Second, the kNN graph construction preceding the GAT layers uses cosine similarity: without normalisation, features with large absolute magnitudes (e.g. byte counts in the millions) would distort the similarity metric and corrupt the attention mechanism's edge weighting [2, 44]. The scaler is fitted on training partition and used identically to development and test

partitions.

3.5.5 Class Balancing via Random Undersampling

Even after five-class consolidation, the BENIGN class remains the numerical majority by a large margin. Random undersampling caps each class at a maximum $N_{\max}$ samples [22, 39]:

$$\hat{n}_c = \min\big(n_c,\, N_{\max}\big) \tag{3.7}$$

This brings the effective IR from $\gg 1000{:}1$ to **1**:**1** and also keeps per-epoch computation tractable, since the unfiltered BENIGN class alone exceeds 2 million samples. After undersampling, the data is shuffled (fixed seed) and split into the final feature matrix $\mathbf{X} \in \mathbb{R}^{N\times 73}$ and label vector $\mathbf{y} \in \{0,1,2,3,4\}^N$.

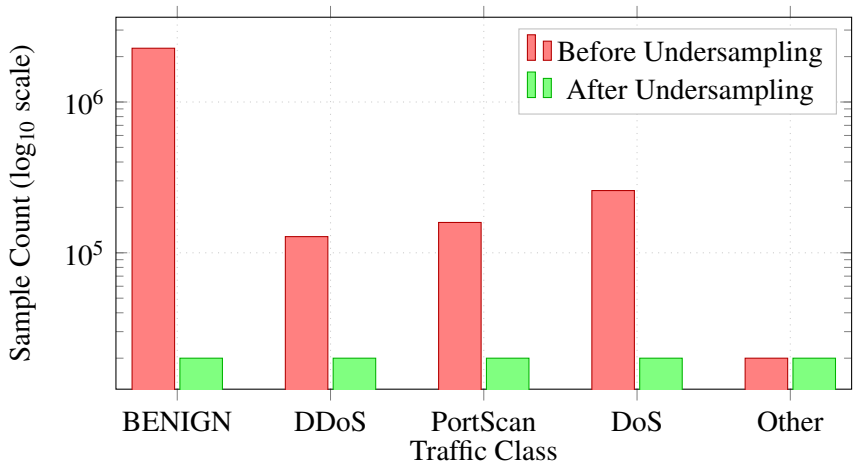


Figure 3.3: Log-scale class distribution before and after random undersampling

CHAPTER 4

PROPOSED METHODOLOGIES

This chapter describes two deep learning systems built to detect cyber threats in different domains: cyberbullying in social media memes, and intrusion attacks in network traffic. The two problems are quite different one involves multimodal content where harmful intent spans both image and text, the other involves flow-level network data where individual flows only reveal their nature in relation to neighbouring flows so each required a different design.

Section 4.1 covers the multimodal cyberbullying detector. It runs parallel image and text branches (DR-HRNet [48] and MC-DeBERTa [49]), fuses their outputs with Multimodal Compact Bilinear (MCB) Pooling, then classifies using Adaptive EfficientNetV2 with a Transformer Spatial Pooling (TSP) module.

Section 4.2 covers the graph-based intrusion detection system. It builds a flow-similarity graph, learns spatial relationships with Graph Attention Networks (GAT) [11], models temporal dynamics with a Bidirectional LSTM (BiLSTM), and outputs five-class traffic labels via an MLP [50], with class-frequency weighted focal loss.

Table 4.1 compares both architectures at a glance.

4.1 Methodology 1 - Multimodal Cyberbullying Detection

4.1.1 Framework Architecture Overview

Meme-based cyberbullying is inherently multimodal: the image alone is often innocent; the offence emerges from the image and caption together. Image-only or text-only classifiers cannot handle this, so the framework runs two dedicated feature-extraction branches in parallel, one per modality, and fuses their outputs before classification.

The pipeline works as follows. A raw meme image J is preprocessed and fed into DR-HRNet to produce a spatial feature map $\mathbf{F}_{\text{img}}$. The corresponding caption T

Table 4.1: Side-by-side comparison of the two proposed architectures.

Component	Cyberbullying	NIDS
Input	Meme image + text caption	Network flow feature vector
Feature Extractor	DR-HRNet (image), MC-DeBERTa (text)	kNN graph + GAT layers
Sequence Model	-	BiLSTM
Fusion	MCB Pooling	Feature concatenation
Classifier	Adaptive EfficientNetV2 + TSP	MLP + Softmax
Loss Function	Binary cross-entropy	Class-frequency weighted Focal loss
Output Classes	2 (Bully / Non-Bully)	5 (BENIGN, DDoS, PortScan, DoS, Other)
Reported Accuracy	99.50%	94.20%

is tokenised and encoded by MC-DeBERTa to give $\mathbf{F}_{\text{txt}}$. MCB Pooling combines both into a joint vector $\mathbf{f}_{\text{fusion}}$, which an Adaptive EfficientNetV2 [51] with a TSP module and softmax head then classifies. Figure 4.1 shows the full block diagram.

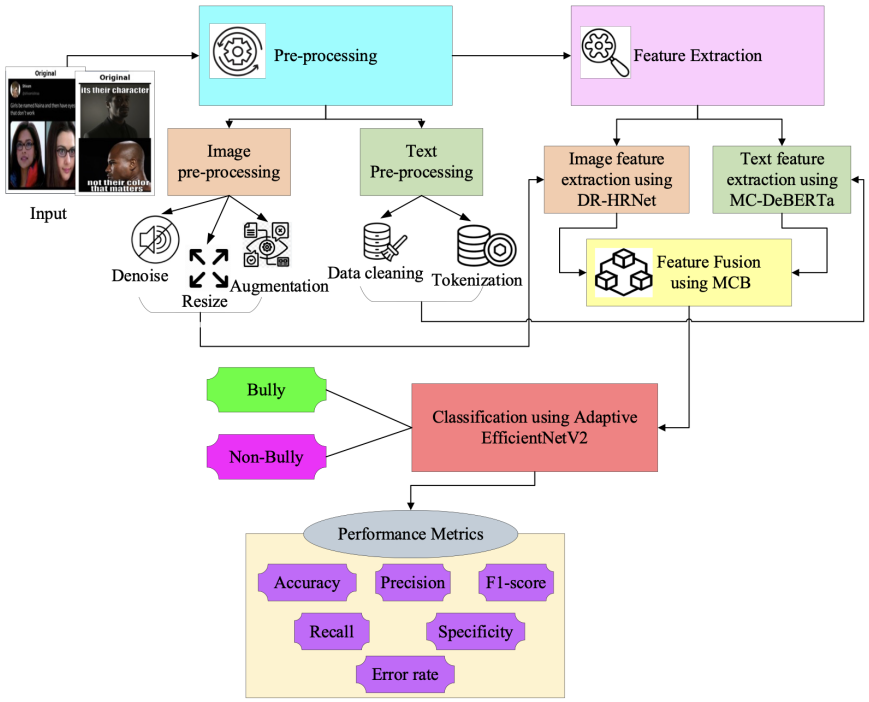


Figure 4.1: Block diagram of the proposed Cyberbully detection model

4.1.2 Image Feature Extraction using DR-HRNet

HRNet Parallel Branch Architecture

After preprocessing, the features of images are obtained using a DR-HRNet [48]. The HRNet design is also appropriate for feature extraction operations where the spatial information is needed because it is designed to support high-resolution representations in the entire network instead of restoring them at subsequent stages. The best way to leverage this ability is to simply eliminate the last pooling layer of the standard HRNet and use the raw high-resolution feature maps that the network has generated directly as the input of further processing. DR-HRNet is designed as shown in Figure 4.2.

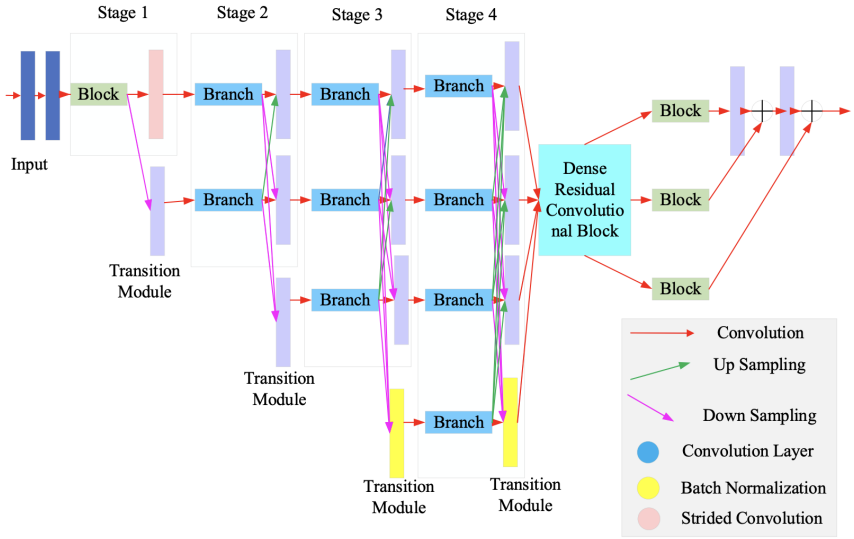


Figure 4.2: DR-HRNet four-stage parallel branch architecture

The HRNet architecture does this using parallel convolutional branches at multiple resolutions with high-to-low resolution feature streams linked in parallel. These parallel branches are repeatedly fused at different resolution levels, allowing the network to constantly communicate between the fine and coarse feature representations to come up with strong, high-resolution feature maps with contextual information. The obtained output feature maps are equal to the sum of the correspondingly up-sampled representations of all the parallel branches, which can be represented as in equation 4.1,

$$\check{f}_S^{(q)} = \sum_{j=1}^{Q} T_{j \to q}\left(f_S^{(q)}\right) \tag{4.1}$$

Here, $f_S^{(q)}$ represents the feature map of the q-th branch, $T_{j \to q}$ represents the resolution

transformation operation, and Q represents the number of parallel branches.

The network takes an input image that has the size of 3×224×2243 and is first filtered by two 3×3 convolutional layers at a 2-stride rate, with the resulting feature maps of size 64×56×5664. HRNet is made of four steps, each of which is a subnetwork comprising several parallel branches. The number of channels and the spatial resolution of each branch are half and double that of the previous branch. Every branch is made with residual blocks, and a block consists of two 3x3 convolutional layers. The initial stage is a high-resolution subnetwork and produces feature maps at the original resolution. Since the second stage, more lower-resolution branches are added gradually, and the high-resolution branches are maintained and continued in all the later stages. Between consecutive stages, fusion layers can be used to allow feature aggregation of resolutions either by upsampling or downsampling to provide dimensional consistency.

Following the fourth step in the HRNet, the multi-resolution fused feature maps are then fed into a DRCB. DRCB combines heavy connectivity with residual learning to enhance feature reuse, contextual representation, and gradient propagation. Using the summation of hierarchical features of the last stage of the HRNet, the DRCB uses the representations obtained to refine them and increase discrimination at multiple scales. Allow the stimulus to the DRCB to be represented as Y_0. The dense version of feature extraction is as in Equation 4.2,

$$F^{\text{final}} = Y_0 + \sum_{l=1}^{L} Y_l \tag{4.2}$$

Here, Y_l represents the output of the l-th convolution layer, and L represents the number of layers in the DRCB. Consequently, the proposed DR-HRNet provides an effective way to combine the spatial preservation capability of HRNet with the strong representational power of dense residual learning, resulting in robust and context-aware feature maps that can be used to achieve high-quality downstream classification.

4.1.3 Text Feature Extraction using MC-DeBERTa

Token, Type and Relative Position Embeddings

MC-DeBERTa is a DeBERTa-v3 encoder extended with a Multi-Scale Dilated Convolutional Block (MDCB). Each token w_j in the input sequence $W = \{w_1, \ldots, w_n\}$

is embedded as:

$$e_j = e^t_j + e^p_j + e^R_j \tag{4.3}$$

where e^t_j is the token embedding, e^p_j is segment type embedding, and e^R_j is relative position embedding. The relative formulation encodes inter-token distances rather than absolute positions, which suits the variable lengths and informal grammar of meme captions.

Disentangled Self-Attention Mechanism

Each Transformer layer uses disentangled self-attention: every token maintains distinct position vectors and content. The attention weights come from 4 interaction terms. The contextualised encoder output is:

$$C_h = \mathrm{Tr}(e) \tag{4.4}$$

where $\mathrm{Tr}(\cdot)$ is the disentangled Transformer encoder and $C_h \in \mathbb{R}^{n\times d}$.

Multi-Scale Dilated Convolutional Block (MDCB)

On top of the Transformer's hidden states, the MDCB applies parallel 1-D dilated convolutions at several dilation rates [49]. Smaller rates capture local n-gram patterns; larger rates capture phrase-level dependencies. For dilation rate D:

$$C_{hD} = \sigma\big(CW_D * C_h + B_D\big) \tag{4.5}$$

All outputs are concatenated and projected to give $\mathbf{F}_{\mathrm{txt}}$. Figure 4.3 shows the full MC-DeBERTa architecture.

4.1.4 Multimodal Feature Fusion using MCB Pooling

MCB is used to combine the extracted image and text features by effectively integrating heterogeneous modalities by modeling the higher-order interaction between visual and textual features. Use the image feature vector f_{image} and the text feature vector f_{text}, the fused representation is calculated as in Equation 4.6,

$$f_{fusion} = \mathrm{MCB}(f_{image}, f_{text}) \tag{4.6}$$

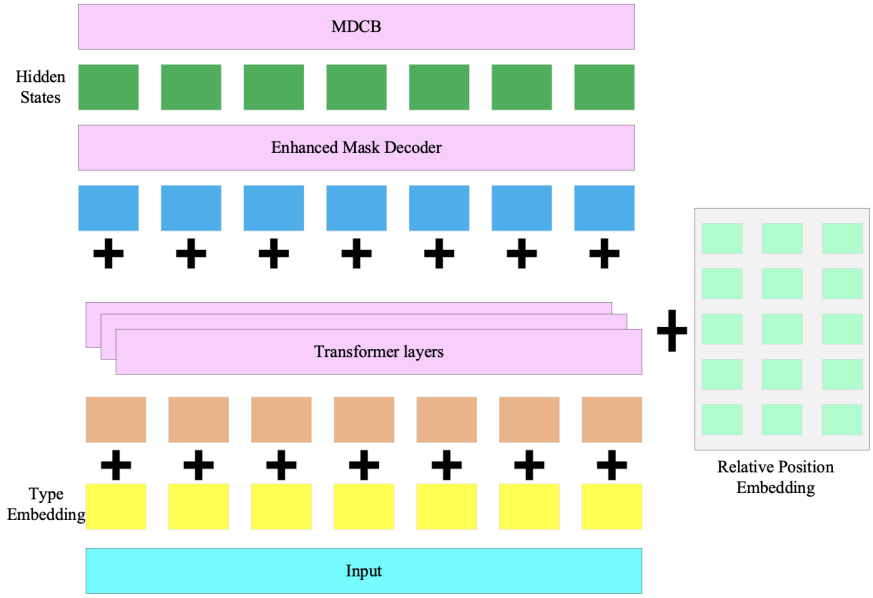


Figure 4.3: MC-DeBERTa architecture

where f_{fusion} represents the joint multimodal feature vector.

4.1.5 Classification using Adaptive EfficientNetV2

Fused-MBConv and MBConv Backbone

The fused feature map can be denoted by the following Equation 4.7,

$$f \in \mathbb{R}^{h \times w \times c} \tag{4.7}$$

The encoded feature map obtained after going through the Adaptive EfficientNetV2 backbone is seen in Equation 4.8,

$$f_e = \varepsilon_{\text{effv2}}(f) \tag{4.8}$$

Where, $f_e \in \mathbb{R}^{\acute{h} \times \acute{w} \times \acute{c}}$ represents the deep discriminative features of the last stage of EfficientNetV2.

Transformer Spatial Pooling (TSP) Module

The TSP module treats spatial positions of f_e as tokens and applies scaled dot-product self-attention to weight regions by their global relevance:

$$\mathrm{Att}(q,k,v) = \mathrm{sm}\left(\frac{qk^T}{\sqrt{D}}\right)v \tag{4.9}$$

Softmax Classification Head

The globally pooled TSP output passes through a fully connected layer to logits $\mathbf{o} \in \mathbb{R}^K$ ($K = 2$), then softmax:

$$p(Y = k) = \frac{\exp(o_k)}{\sum_{i=1}^{K} \exp(o_i)} \tag{4.10}$$

The model trains end-to-end with binary cross-entropy loss and Adam. On the Facebook Hateful Meme Dataset it achieves **99.45%** accuracy and **99.30%** F1-score.

4.2 Methodology 2 - Graph-Based Hybrid Network Intrusion Detection

4.2.1 Framework Architecture Overview

In network traffic, a single flow rarely tells you much on its own. A SYN packet is unremarkable; thousands from the same source in a second is a SYN flood. Flat feature-vector models miss this because they treat each flow independently. The proposed framework models flows as nodes in a similarity graph, learns the graph's spatial topology with GAT, captures temporal flow ordering with BiLSTM, and classifies into five traffic types with an MLP.

The preprocessed feature matrix $\mathbf{X} \in \mathbb{R}^{N\times 73}$ feeds a kNN graph construction step, two stacked multi-head GATConv layers, a BiLSTM, and finally an MLP with class-frequency weighted softmax. Figure 4.4 shows the full architecture.

4.2.2 Graph Construction using k-Nearest Neighbours

The feature matrix $\mathbf{X}$ is converted into a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ where each flow is a node. An edge (v_i, v_j) is added when v_j is among the $k = 10$ nearest neighbours of v_i

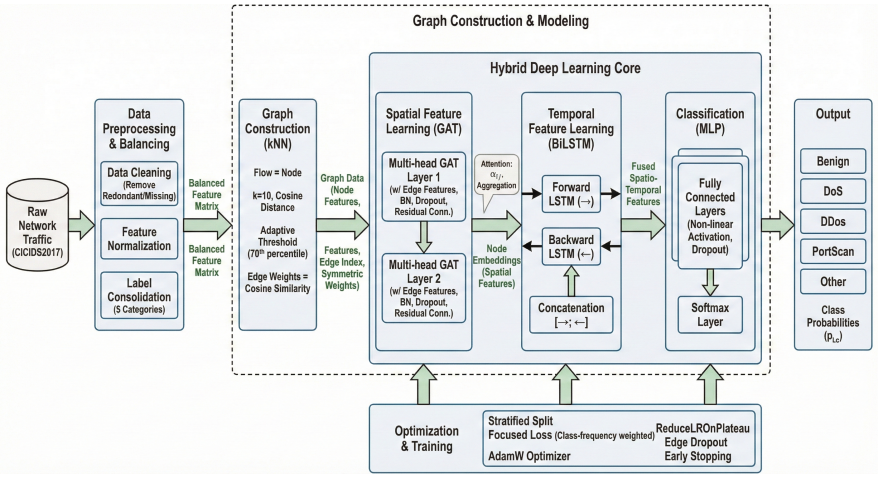


Figure 4.4: Architecture of the proposed hybrid GAT-BiLSTM-MLP intrusion detection framework

by cosine similarity:

$$\cos\_\text{sim}(x_i, x_j) = \frac{x_i^\top x_j}{\|x_i\|_2 \cdot \|x_j\|_2} \tag{4.11}$$

Cosine similarity is preferable to Euclidean distance here because it is magnitude-invariant, which matters given the wide numerical range of standardised flow features. Edges beyond the 70th-percentile distance threshold are pruned to remove spurious long-range connections [41].

Since kNN is directional by nature, the graph is symmetrised: for every directed edge $(v_i \to v_j)$, the reverse $(v_j \to v_i)$ is added with equal weight. Symmetric message passing stabilises the learned node representations.

4.2.3 Spatial Feature Learning using GAT

Standard graph convolutions aggregate all neighbours with equal weight. In network traffic this is wrong: a DDoS flow and a DNS query that happen to be similar in feature space should not contribute equally to each other's representation. GAT learns per-edge attention weights instead [11]. The raw attention score for an edge $j \to i$ is:

$$e_{ij} = \text{LeakyReLU}\Big(\mathbf{a}^\top \big[Wh_i \,\|\, Wh_j\big]\Big) \tag{4.12}$$

normalised across the neighbourhood $\mathcal{N}(i)$:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \tag{4.13}$$

Node representations update via:

$$h_i' = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W h_j \tag{4.14}$$

Multi-head attention with $K = 8$ is used to capture diverse neighbourhood patterns. Heads are concatenated in the first GATConv layer and averaged in the second:

$$h_i^{(1)} = \Big\|_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(k)} W^{(k)} h_j \right) \tag{4.15}$$

Each GATConv layer is followed by dropout, batch normalisation and a residual connection. The resulting node embeddings $\mathbf{H}^{\text{GAT}} \in \mathbb{R}^{N \times d_{\text{GAT}}}$ are passed to the BiLSTM.

4.2.4 Temporal Feature Learning using BiLSTM

The GAT layers encode which flows are structurally similar; the BiLSTM then captures how flows evolve over time. The standard LSTM gates are [11]:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{4.16}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{4.17}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{4.18}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad h_t = o_t \odot \tanh(c_t) \tag{4.19}$$

The BiLSTM reads the GAT output in both directions and concatenates the results, giving each node a representation conditioned on both past and future context:

$$\overrightarrow{h}_t = \text{LSTM}(x_t, \overrightarrow{h}_{t-1}) \tag{4.20}$$

$$\overleftarrow{h}_t = \text{LSTM}(x_t, \overleftarrow{h}_{t+1}) \tag{4.21}$$

$$h_t^{(\text{BiLSTM})} = \left[\overrightarrow{h}_t \, ; \, \overleftarrow{h}_t \right] \tag{4.22}$$

4.2.5 Classification using MLP

The BiLSTM output passes through a two-layer MLP [2]:

$$z_t^{(1)} = \mathrm{ReLU}\big(W_1\, h_t^{(\mathrm{BiLSTM})} + b_1\big) \tag{4.23}$$

$$z_t^{(2)} = \mathrm{ReLU}\big(W_2\, z_t^{(1)} + b_2\big) \tag{4.24}$$

with dropout (0.3) after each ReLU. The five-class prediction is:

$$p_{t,c} = \frac{\exp(o_{t,c})}{\sum_{k=1}^{C} \exp(o_{t,k})}, \qquad \hat{y}_t = \arg\max_c\, p_{t,c} \tag{4.25}$$

To reduce the effect of the remaining class imbalance after undersampling, a class-frequency weighted focal loss was used [22]. The class weight for class c is computed as:

$$\alpha_c = \frac{N}{\hat{n}_c} \tag{4.26}$$

The focal loss function is defined as:

$$\mathscr{L}_{\mathrm{FL}} = -\frac{1}{N} \sum_{t=1}^{N} \alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{4.27}$$

where $\gamma = 2.0$ controls focus on hard-to-classify samples, p_t predicted probability of the true class for sample t, and α_t denotes the weight assigned to the corresponding class.

4.2.6 Training Strategy

Graph nodes are split 70/15/15 using stratified sampling, which preserves class proportions across all three sets. Without this, rare attack classes could end up under-represented in the test set and produce over-optimistic numbers.

The model trains with AdamW, which couples Adam's adaptive updates with decoupled L_2 weight decay:

$$\theta_{t+1} = \theta_t - \eta \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \varepsilon) - \eta \lambda \theta_t \tag{4.28}$$

A ReduceLROnPlateau scheduler drops the learning rate by factor $\rho < 1$ whenever

validation loss stalls, and early stopping restores the lowest-validation-loss checkpoint as the final model. On CICIDS2017 the model achieves **94.20%** accuracy, **94.53%** precision, **94.20%** recall, and **94.20%** F1-score.

4.3 Summary

For **cyberbullying,** the system extracts image features via DR-HRNet + DRCB and text features via MC-DeBERTa + MDCB (Eq. 4.3–4.5), fuses them with MCB Pooling , and classifies with EfficientNetV2 + TSP (Eq. 4.9–4.10), reaching **99.5%** accuracy on the Facebook Hateful Meme Dataset [47].

For **NIDS,** the system builds a kNN cosine-similarity graph (Eq. 4.11), learns spatial topology through two GAT layers (Eq. 4.12–4.15), models temporal dynamics with BiLSTM (Eq. 4.16–4.22), and classifies with a weighted-loss MLP (Eq. 4.23–4.27), reaching **94.20%** accuracy on CICIDS2017 [44].

The experimental evaluation for both models is in Chapter 5.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Implementation Details

5.1.1 Training Configuration and Hyperparameters

Both frameworks were built in Python 3.10 and trained using PyTorch with GPU acceleration. The cyberbullying model and the intrusion detection model have different architectural constraints, so their training setups differ in a few meaningful ways.

Hyperparameter Settings

For the cyberbullying detection framework, the simulation parameters used during training are listed in Table 5.1.

Table 5.1: Training hyperparameters for the cyberbullying detection framework

Parameter	**Value**
Optimizer	Adam
Learning Rate	1×10^{-3}
Loss Function	CrossEntropy
Batch Size	16
Epochs	100
Training Split	80%
Testing Split	20%
Input Image Size	224×224
Text Sequence Length	20 tokens (padded/truncated)
Text Embedding Dimension	100
MC-DeBERTa LSTM Hidden Dim	128 (bidirectional)
Image Feature Dim (DR-HRNet)	128
Text Feature Dim (MC-DeBERTa)	128
Fused Feature Dim (MCB)	256
Classifier Hidden Dim	128
Dropout (Classifier)	0.4
DataLoader Shuffle	True

For the network intrusion detection framework, the hyperparameters are listed in Table 5.2.

Table 5.2: Training hyperparameters for the network intrusion detection framework

Parameter	**Value**
Optimizer	AdamW
Learning Rate	8×10^{-4}
Weight Decay	1×10^{-5}
Betas	$(0.9,\ 0.999)$
Loss Function	Focal Loss ($\gamma = 2.0$, class-frequency α)
Hidden Feature Dim	128
GAT Attention Heads	8
BiLSTM Hidden Size	64 (per direction)
Dropout Rate	0.3
Edge Dropout Rate	0.1
Gradient Clipping	max_norm = 1.0
Max Epochs	50
Early Stopping Patience	25 epochs
LR Scheduler	ReduceLROnPlateau (factor = 0.5, patience = 5)
Training Split	70%
Validation Split	10%
Testing Split	20%
kNN Graph (k)	10 (cosine distance, 70th percentile threshold)

Optimisation Techniques

For the cyberbullying model, the Adam optimiser with CrossEntropy loss handled the binary classification across the two modalities.

For the intrusion detection model, a class-frequency-based Focal Loss (gamma = 2.0) was used to address the heavy class imbalance in network traffic. This loss function up-weights minority attack classes during backpropagation so the model does not simply overfit to BENIGN traffic. A `ReduceLROnPlateau` scheduler reduced the learning rate by a factor of 0.5 when validation accuracy stalled for five consecutive epochs. Training ran for a maximum of 50 epochs, with early stopping triggered if no improvement in validation accuracy was observed for 25 consecutive epochs. The split was 70% training, 10% validation, and 20% testing, all stratified by class.

5.2 Results and Discussion – Cyberbullying Detection

The cyberbullying detection framework was evaluated on the Facebook Hateful Meme Dataset, that have both image and text data across Bully and Non-Bully classes. Dataset has 10,000 text samples. Dataset contained 37.7% bully samples and 62.3% non-bully samples.

5.2.1 Training and Testing Accuracy and Loss

Training Accuracy Curve

FHM curve is noisy but it climbs from 0.4 to near perfect by epoch 100, which is model actually learning.

Validation Accuracy Curve

The validation accuracy curve tracks the training curve. It converging near 1.0 by the end of training. The two curves do not change drastically, which indicates the model is generalising. The smooth convergence relative to Dataset reflects the more balanced class distribution in the Hateful Meme Dataset.

Training Loss Curve

Training loss drops sharply in the first 20 epochs and continues to decrease through the full 100 epochs. The rapid descent reflects the effectiveness of the multi-modal fusion in quickly capturing cross-modal dependencies between image and text features.

Validation Loss Curve

The validation loss tracks the training loss closely and levels off after the first few epochs. The two curves stay close throughout, with no meaningful divergence between them.

The training and testing accuracy and loss curves are represented in Figure 5.1.

5.2.2 Confusion Matrix Analysis

The confusion matrix for the Facebook Hateful Meme Dataset is shown in Figure 5.2.

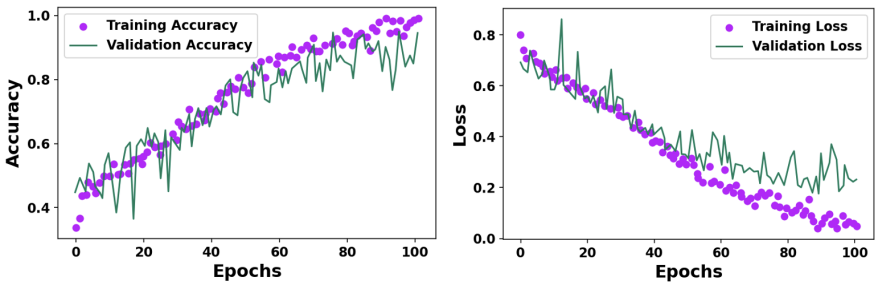


Figure 5.1: Training and Testing Accuracy and Loss Curve

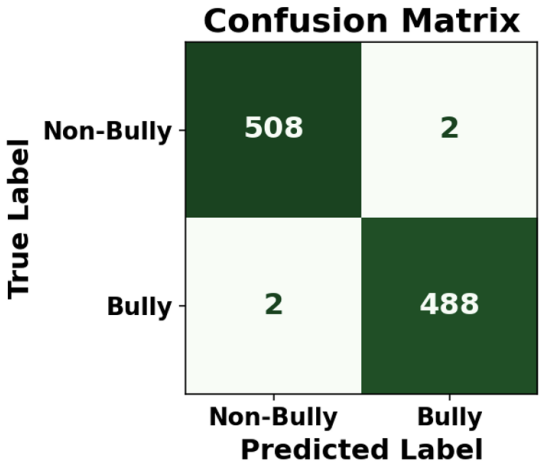


Figure 5.2: Confusion matrix of Cyberbully Detection

Bully Class Detection Performance

Out of 490 occurrences of Bully in the actual dataset, the algorithm successfully classified 488 occurrences, and misclassified just 2. Thus, the false negative classification rate for Bully is below 0.5%. The importance of an exceptionally low false negative classification rate for Bully is due to the good performance of the algorithm in this area.

Non-Bully Class Detection Performance

Out of 510 actual Non-Bully samples, 508 were correctly classified and 2 were incorrectly flagged as Bully. A false positive rate this low means the system would not generate excessive spurious alerts in practice. The near-symmetry between false positives and false negatives across both classes reflects a well-calibrated classifier that is not trading one type of error for the other.

5.2.3 ROC Curve and AUC Analysis

ROC Curve Interpretation

The ROC curve for the FHM Dataset shows both the Bully and Non-Bully class curves pushing very close to the top-left corner. This shape means model achieves high tpr before accumulating any meaningful number of false positives the ideal behaviour for a detection system.

AUC Score Analysis

The AUC scores on the Facebook Hateful Meme Dataset are 99.21% for Bully class and 99.34% for Non-Bully class. Both values are very close to 1.0, confirming that model's ability to separate 2 classes is robust across all possible decision, not just the one chosen during evaluation.

Figure 5.3 demonstrates the ROC curve.

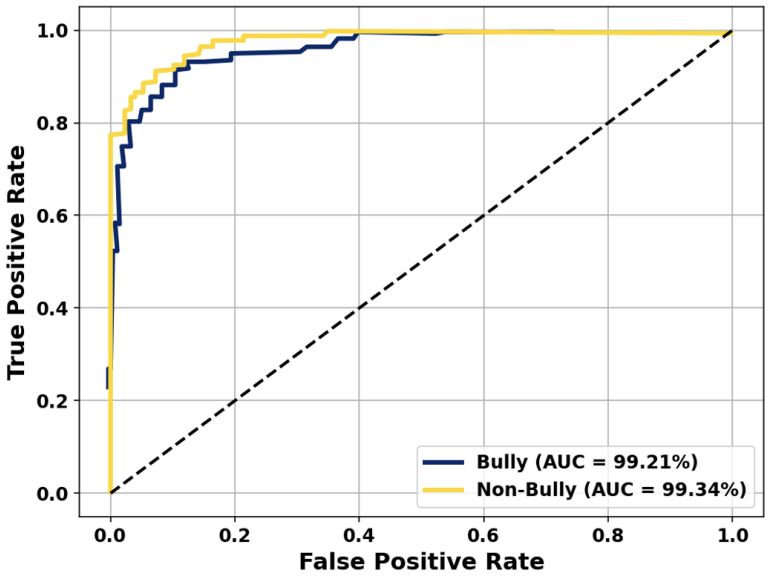


Figure 5.3: ROC curve of Cyberbully Detection

5.2.4 Performance Comparison Metrics

The proposed DR-HRNet + MC-DeBERTa + Adaptive EfficientNetV2 framework is compared against five baseline models on the FHM Dataset.

Accuracy Comparison

The model achieves 99.5% accuracy on Bully and 99.2% on Non-Bully classes, well above MobileNetV2 (the next best at 94.00% and 93.00% respectively). The gap is large enough that it cannot be attributed to random variation. Figure 5.4 represents the Accuracy.

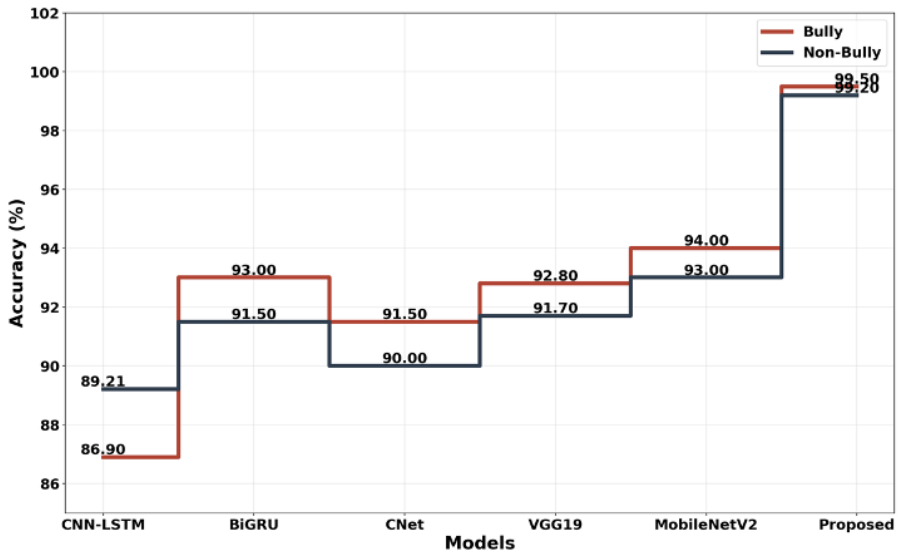


Figure 5.4: Accuracy of Cyberbully Detection

Precision Comparison

The proposed model achieves 99.6% precision on Bully and 99.3% on Non-Bully classes. Figure 5.5 represents the Precision.

Recall Comparison

The proposed model achieves 99.1% recall on Bully and 99.4% on Non-Bully classes. Figure 5.6 represents the Recall.

F1-Score Comparison

The proposed model achieves 99.5% recall on Bully and 99.2% on Non-Bully classes. Figure 5.7 represents F1 Score.

In all four metrics, the proposed model consistently outperforms every baseline on the FHM Dataset. Nearest competitor (MobileNetV2) trails by several percentage points on every metric.

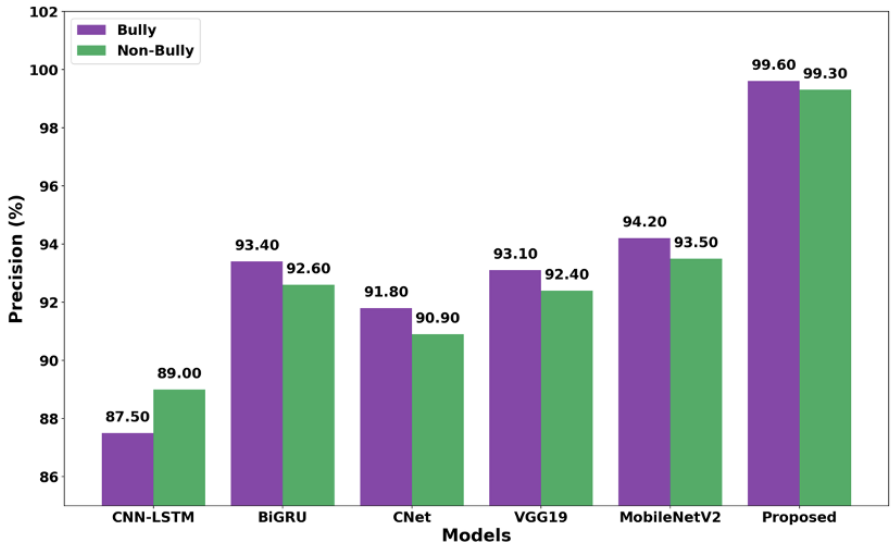


Figure 5.5: Precision of Cyberbully Detection

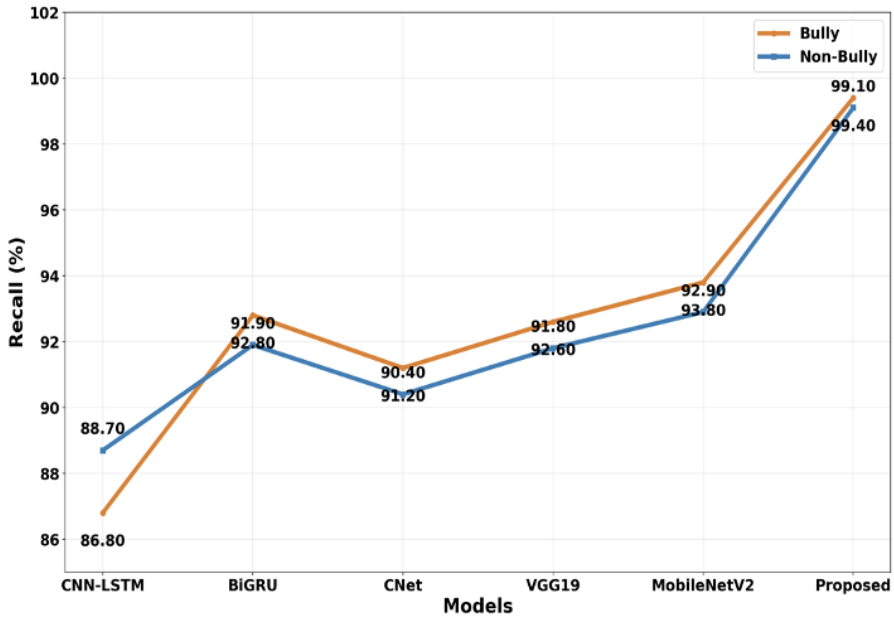


Figure 5.6: Recall of Cyberbully Detection

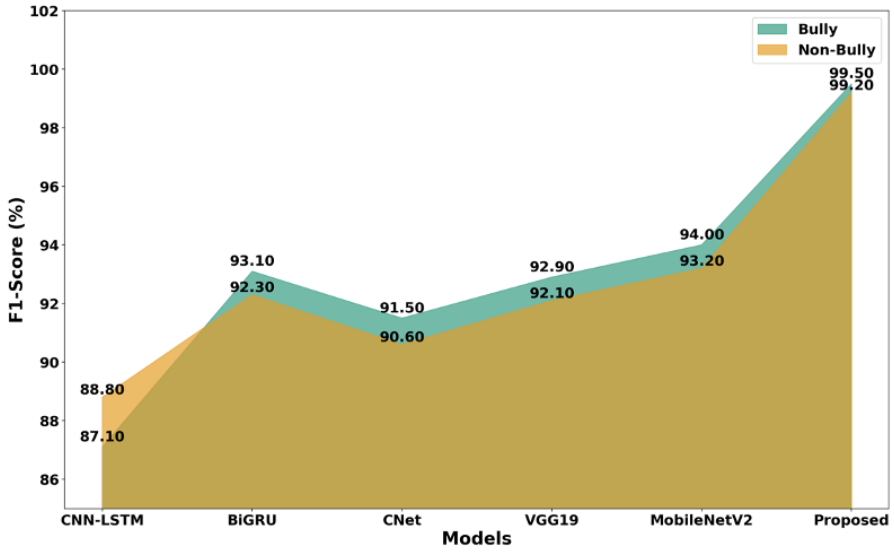


Figure 5.7: F1 Score of Cyberbully Detection

5.2.5 Discussion of Results

Impact of DR-HRNet

DR-HRNet handles image feature extraction by preserving multi-scale spatial representations throughout the network. CNNs that downsample aggressively early on, DR-HRNet maintains high-resolution branches in parallel and fuses them using dense residual connections. For FHM Dataset, the hateful intent is frequently expressed in the visual. In the case of cues, this spatial fidelity is important. The model can detect texture, composition and The visual elements that would not be captured without higher resolution representation, in the context of which they are used.

Impact of MC-DeBERTa

MC-DeBERTa is extension of DeBERTa architecture that extracts text features. To get multi-scale contextual information. Using attention over content and relative vocabulary. It is capable of processing informal language, online slang, and implicit references, like position embeddings. Much improved over previous models. Captions of hateful memes usually consist of few words, but they are packed with context. MC-DeBERTa's ability to recognize relationship between positions in short sequences allows it to do so. Suitable text for this type of text.

Effectiveness of MCB Fusion

Multimodal Compact Bilinear Pooling combines the image and text feature vectors into joint representation that represents cross-modal interactions. MCB models the pairwise correlations between image and text features without memory explosion that full bilinear pooling would require. For FHM Dataset, where hateful content often only makes sense when image and caption are considered together, this fusion step is what ties two modalities into coherent representation.

Performance of Adaptive EfficientNetV2

Adaptive EfficientNetV2, augmented with a Transformer Spatial Pooling (TSP) module, performs the final classification over the fused multimodal features. The TSP module allows the classifier to attend to spatially and semantically relevant regions of the fused representation rather than treating all positions equally. The very high accuracy (99.50%) and AUC scores (99.21% and 99.34%) on the Facebook Hateful Meme Dataset suggest that this combination MCB fusion feeding into a spatially-aware classifier is well-suited to the multimodal nature of the problem.

5.3 Results and Discussion - Network Intrusion Detection

The framework was evaluated on the CICIDS2017 dataset, which contains flow-based records of both benign and attack traffic. After preprocessing and label consolidation, five traffic categories were used: BENIGN, DDoS, DoS, PortScan, and Other. Class balancing via random undersampling was applied to prevent the model from favouring to majority-class predictions.

5.3.1 Confusion Matrix Analysis

The confusion matrix for the Hybrid GAT+BiLSTM+MLP model is reproduced in Table 5.3.

Table 5.3: Confusion matrix - Hybrid GAT+BiLSTM+MLP (CICIDS2017)

True \Predicted	BENIGN	DDoS	DoS	Other	PortScan
BENIGN	2533	111	11	2	2
DDoS	1	2608	50	0	0
DoS	29	224	2372	33	1
Other	12	9	2	2373	263
PortScan	0	0	8	13	2638

BENIGN Traffic Classification

Of 2,659 BENIGN samples, 2,533 were correctly classified. The main source of misclassification was DDoS (111 samples), likely because high-volume benign traffic can share statistical properties with DDoS flows. The false positive rate for BENIGN is nonetheless low relative to the class size.

DDoS Attack Classification

DDoS classification is strong, with 2,608 out of 2,659 samples correctly identified. Only 1 DDoS sample was misclassified as BENIGN, which is exactly the kind of error that matters most in a real IDS. The 50 samples confused with DoS are understandable - both attack types involve traffic volume patterns that can overlap.

DoS Attack Classification

DoS shows the highest within-class confusion, with 224 samples misclassified as DDoS and 29 as BENIGN. This is expected given the structural similarity between DoS and DDoS flows when analysed at the flow-feature level. Despite this, 2,372 out of 2,659 samples are correctly identified.

PortScan Attack Classification

PortScan achieves near-perfect classification, with 2,638 out of 2,659 samples correct and only 8 and 13 samples confused with DoS and Other respectively. PortScan has a distinctive low-bandwidth, high-connection-count signature that the graph-based model captures well.

Other Attack Classification

The Other class, which aggregates less frequent attack types, is the most challenging category. 2,373 samples are correctly identified, but 263 are confused with PortScan. This confusion is partly a consequence of the label grouping - the Other category contains heterogeneous attacks whose flow-level features may partially resemble PortScan patterns.

Figure 5.8 represents confusion matrix.

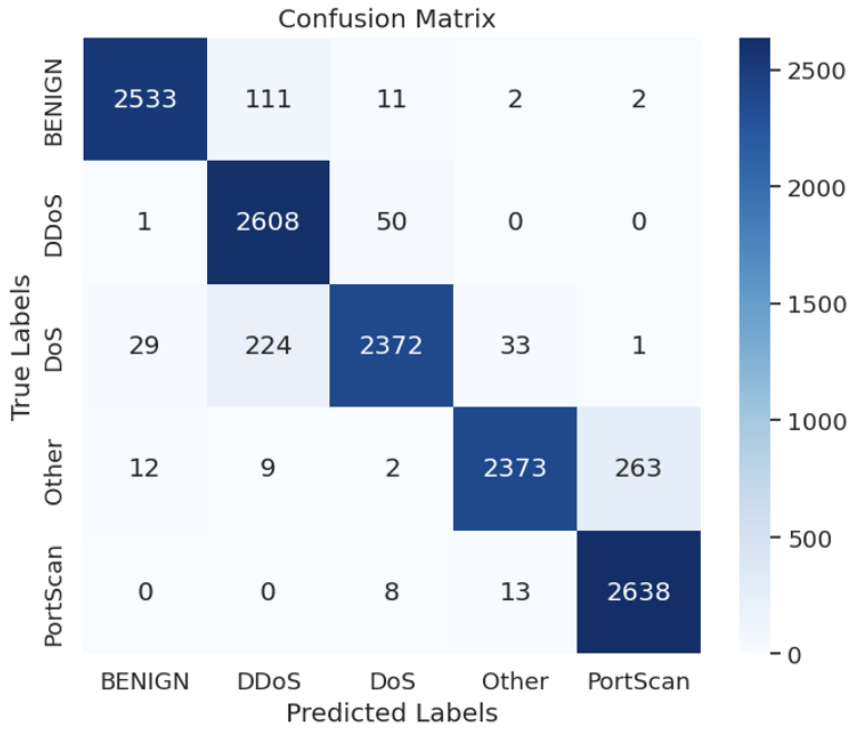


Figure 5.8: Confusion Matrix (Proposed model)

5.3.2 ROC Curve and AUC Analysis

Multi-Class ROC Curve

The multi-class ROC curves for the Hybrid GAT+BiLSTM+MLP model show all five class curves pushed tightly toward the top-left corner of the plot. Each class curve rises steeply before any meaningful accumulation of false positives - the hallmark of a classifier with strong discriminative ability. This behaviour holds across all five traffic categories, including the harder-to-separate Other class.

Macro-Average AUC Analysis

The per-class AUC scores for the Hybrid GAT+BiLSTM+MLP model are:

All five classes exceed 0.98, and the macro-average AUC is 0.997. This means that regardless of which traffic class is being evaluated, the model can reliably separate it from the rest. The DoS class has the lowest individual AUC (0.988), consistent with the confusion matrix showing some DoS/DDoS overlap.

Figure 5.9 represents confusion matrix.

Table 5.4: Per-class AUC scores of Hybrid GAT+BiLSTM+MLP

Traffic Class	AUC
BENIGN	0.998
DDoS	0.997
DoS	0.988
Other	0.994
PortScan	0.998

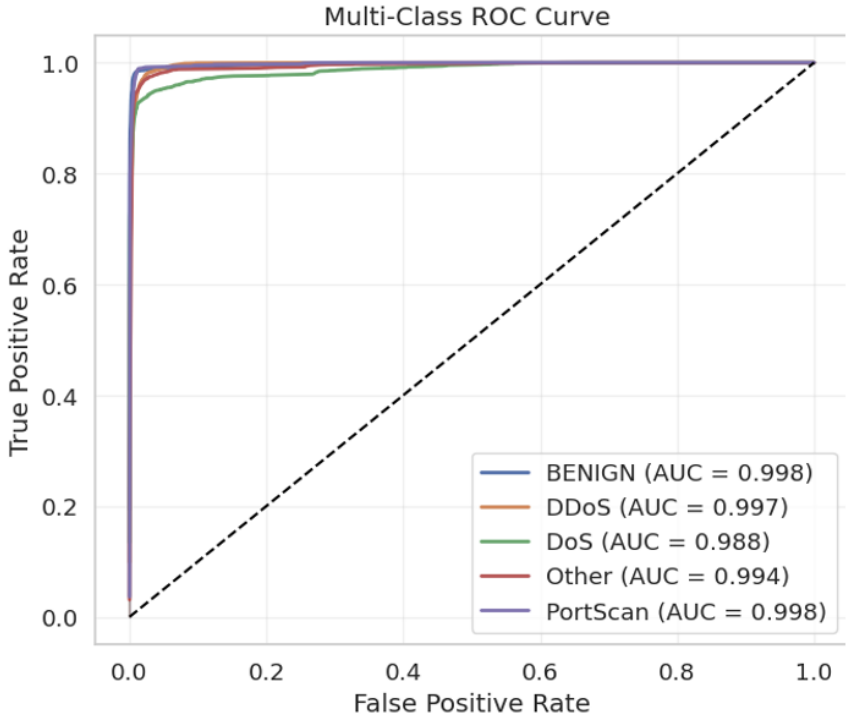


Figure 5.9: ROC Curve (Proposed model)

5.3.3 Performance Metrics Evaluation

Accuracy Evaluation

The proposed model achieves an accuracy of 94.20%. Given the five class structure the inherent difficulty of separating correlated attack types, this is a strong result. For a deployed IDS, high precision is important to avoid alert fatigue due to excessive. False positives harm analysis effectiveness.

Precision Evaluation

Overall precision is 94.53%, meaning that when the model raises an alert for a particular traffic class, it is correct more than 94% of the time. For a deployed IDS, high precision is important - alert fatigue from excessive false positives degrades analyst effectiveness.

Recall Evaluation

Due to matching accuracy figure, recall of 94.20% indicates that the model is not missing attacks in a systematic manner. The balance of precision and recall is a direct result of the focal loss and class balancing strategy.

F1-Score Evaluation

The value of 94.20% F1-score confirms the model is not gaming either metrics at the cost of the other. On an imbalanced five class data set , macro F1 that is this high indicates that model performing consistently across the majority and minority classes.

5.3.4 Comparison with Baseline Models

Table 5.5 and Figure 5.10 represents the full performance comparison.

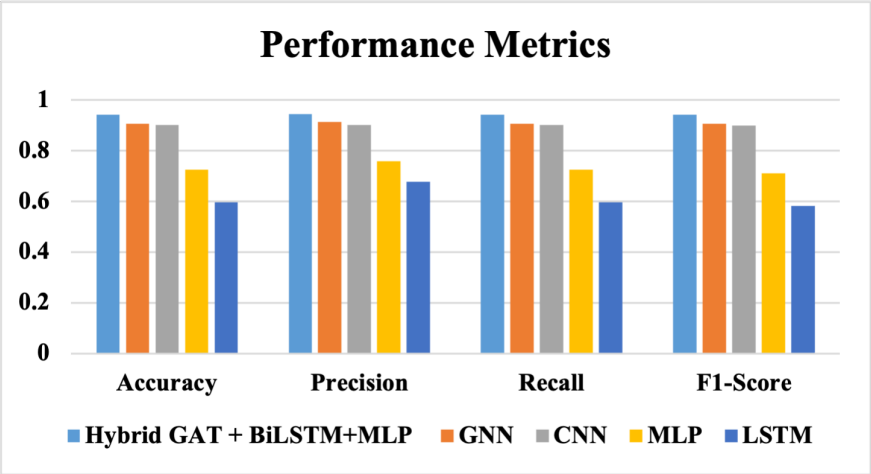


Figure 5.10: Performance Metrics accross models on NIDS

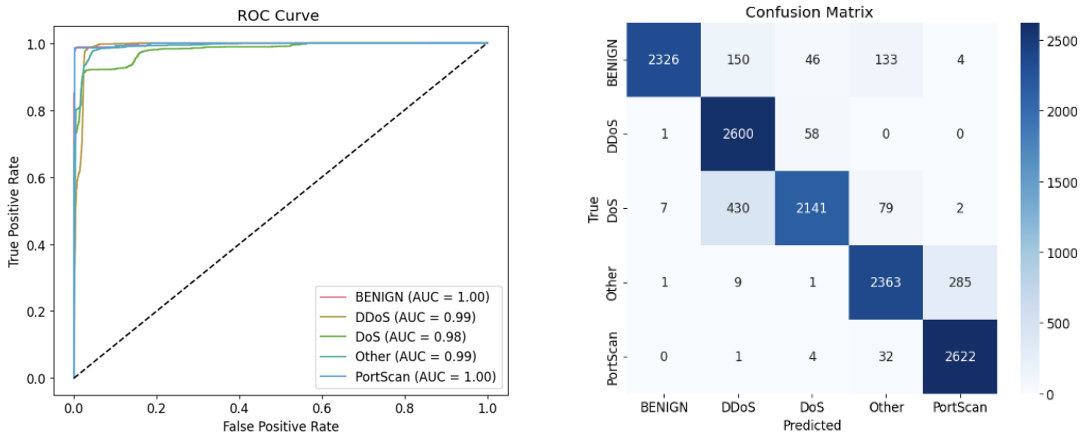
Comparison with GNN

The standalone GNN demonstrates an impressive accuracy of 0.9065, and the macro AUC across classes is competitive we can see from the per class ROC curves that it is effecting strong class separation. The accuracy and F1 of the hybrid is beaten by it

Table 5.5: Performance comparison across models on NIDS

Model	Accuracy	Precision	Recall	F1-Score
Hybrid GAT + BiLSTM + MLP	**0.9420**	**0.9453**	**0.9420**	**0.9420**
GNN	0.9065	0.9139	0.9065	0.9063
CNN	0.9013	0.9023	0.9013	0.9004
MLP	0.7258	0.7585	0.7258	0.7107
LSTM	0.5953	0.6774	0.5953	0.5830

by some 3.5 percentage points. The reason behind this gap is that the local topological relationships of the flows are represented by the GNN, but not the sequential dependence that the BiLSTM does. This difference is most visible in multi-stage attacks spread over time. Figure 5.11 represents GNN confusion matrix and ROC curve.



(a) ROC Curve (b) Confusion Matrix

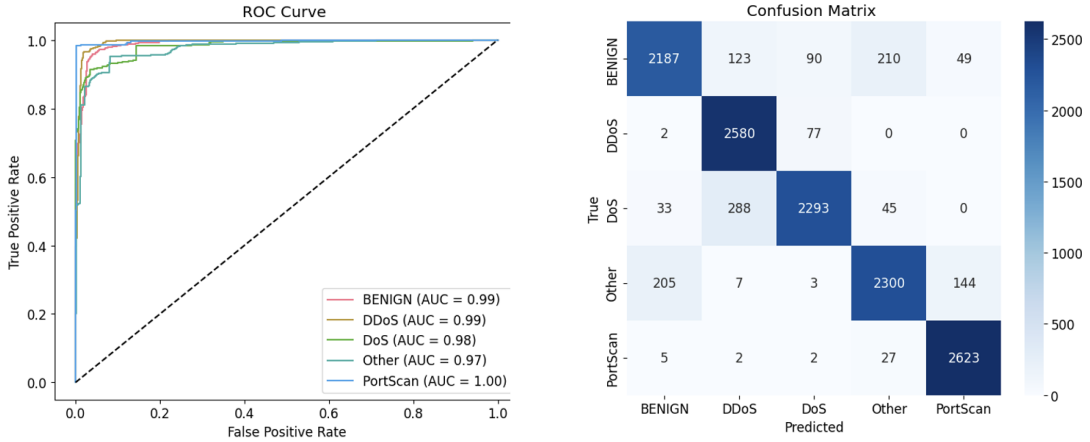
Figure 5.11: Performance evaluation of the GNN model.

Comparison with CNN

The CNN model reaches 90.13% accuracy, just below the GNN. CNNs can learn local feature patterns from flow vectors, but they do not model the relational structure between flows or temporal dependencies. The confusion matrix for the CNN shows slightly more off-diagonal errors than the GNN, particularly for the Other class. Figure 5.12 represents CNN confusion matrix and ROC curve.

Comparison with MLP

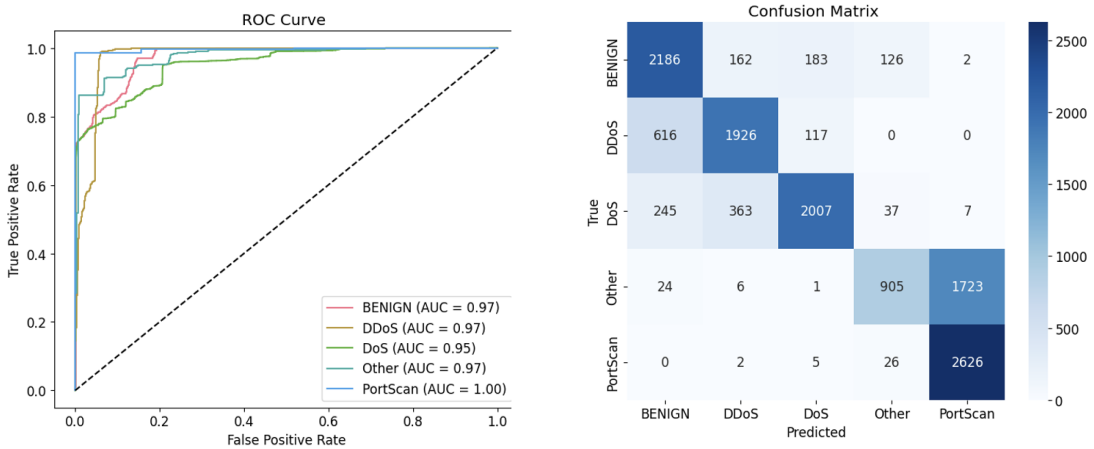
The MLP model drops noticeably, to 72.58% accuracy and 71.07% F1. The confusion matrix shows considerably more misclassifications across all classes compared to the graph-based models. This confirms that treating each network flow as



(a) ROC Curve (b) Confusion Matrix

Figure 5.12: Performance evaluation of the CNN model.

an independent input, without any relational or sequential context, loses too much structural information for reliable multi-class intrusion detection. Figure 5.13 represents MLP confusion matrix and ROC curve.



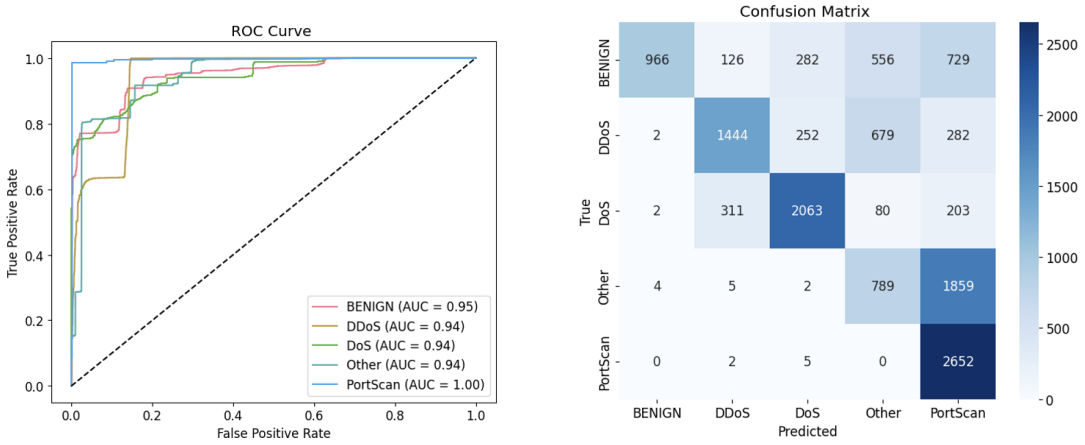
(a) ROC Curve (b) Confusion Matrix

Figure 5.13: Performance evaluation of the MLP model.

Comparison with LSTM

The standalone LSTM is the weakest baseline at 59.53% accuracy and 58.30% F1. Its ROC curves show noticeably lower AUC values, particularly for BENIGN (0.95) and DDoS/DoS (0.94), compared to the hybrid model. Sequence modelling alone, without the spatial graph structure that GAT provides, is not enough to reliably classify the five traffic categories. The LSTM results effectively show what happens when relational structure is ignored and the answer is a substantial performance drop. Figure 5.14

represents CNN confusion matrix and ROC curve.



(a) ROC Curve (b) Confusion Matrix

Figure 5.14: Performance evaluation of the LSTM model.

5.3.5 Discussion of Results

Impact of Graph Attention Learning

The GAT component does more than aggregate neighbourhood information - it selectively weights neighbours based on their relevance to each node. For DDoS and PortScan, which involve coordinated flows with distinctive interaction patterns, this attention mechanism groups similar flows together in the feature space. The confusion matrix confirms this: DDoS and PortScan have the sharpest on-diagonal concentrations in the proposed model's matrix compared to all baselines.

Effectiveness of BiLSTM Temporal Modelling

The performance gap between the GNN and the full hybrid model directly quantifies what BiLSTM adds. Complex attacks like DoS variants and multi-stage intrusions play out across a sequence of flows - the GNN sees relationships at a point in time, but the BiLSTM captures how those relationships evolve. The bidirectional processing is particularly useful for attacks that have both a build-up and a wind-down phase in the traffic sequence.

Contribution of MLP Classification

The MLP classifier at the end of the pipeline takes the spatially and temporally enriched node embeddings from GAT and BiLSTM and maps them to class probabilities. Its role is to learn the non-linear decision boundaries that separate the five traffic

categories once the upstream components have done the feature work. combination of focal loss and class-weighted training pushes MLP to perform well on minority attack classes (Other, DoS) rather than just on majority class.

Effect of Class Balancing Strategy

CICIDS2017 is dominated by BENIGN traffic. A naive model would score reasonably on accuracy just by predicting BENIGN most of the time, while failing badly on recall for minority attack classes. Random undersampling, combined with class-frequency-based focal loss, forces the model to treat all five categories seriously. The near-equal precision and recall figures (94.53% and 94.20%) across such different class types show that this strategy worked as intended.

5.4 Comparative Analysis Across Domains

5.4.1 Comparative Performance Analysis

Table 5.6 summarises the best achieved metrics across both domains.

Table 5.6: Cross-domain performance summary

Domain	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Cyberbullying Detection	99.50	99.60	96.40	99.50
Network Intrusion Detection	94.20	94.53	94.20	94.20

The cyberbullying model achieves higher raw numbers across the board. This is expected: the Facebook Hateful Meme Dataset is a binary classification problem on a relatively balanced two-class split. The intrusion detection problem is five-class with structurally similar attack categories, which is inherently harder. Still, 94.20% macro F1 on a five-class imbalanced dataset is a strong result by the standards of the field.

5.4.2 Deep Learning Effectiveness Across Domains

Both domains benefit from deep learning in different ways. For cyberbullying detection, the advantage comes from multimodal representation no single-modality baseline comes close to the proposed framework's performance, which is direct evidence that combining image and text features captures information that neither alone provides. For intrusion detection, the advantage comes from structural and temporal representation the GNN and CNN baselines show that graph features alone or local patterns alone are

insufficient; both spatial topology and sequential dynamics are needed.

Transformer-based components (MC-DeBERTa, TSP in EfficientNetV2) drive the text and multimodal classification gains in Cyberbully Detection. Graph attention and bidirectional sequence modelling drive the gains in Network Intrusion Detection. The underlying principle is the same in both cases: architecture choices should match the structure of the data.

5.4.3 Challenges in Multimodal and Network Data

Multimodal cyberbullying data is difficult because the label is often a function of both modalities simultaneously. A caption that looks benign and an image that looks benign can become hateful meme content in combination this is the defining challenge of the Hateful Meme Dataset. The MCB fusion is the component that addresses this directly, but it also adds complexity to the training pipeline.

Network traffic data is difficult for different reasons. The class imbalance is severe, the boundary between DoS and DDoS is not crisp at the flow-feature level, and the Other category is inherently heterogeneous. The confusion matrix for the model reflects this - the only notable off-diagonal concentrations are DoS/DDoS and Other/PortScan, both of which correspond to genuinely ambiguous pairs at the feature level rather than model failures.

5.4.4 Strengths and Limitations of the Proposed Frameworks

Both frameworks yield significant accomplishment on their tasks and outperform their baselines by significant margins. Compared to missing person detection and similar problems, cyberbullying is relatively cleaner two class problem with more or less balanced data. With score of over 0.99, test set is less subject to scepticism. With respect to intrusion detection aspect, class imbalance very well managed, and consistent performance is observed on all five traffic categories.

The cyberbullying framework was tested on a specific dataset and has not been evaluated on live social media streams, where preprocessing assumptions may not hold and class balance may shift substantially. The intrusion detection model was built on CI-CIDS2017, which, despite being widely used, has known limitations its simulated attack traffic may not capture full diversity of real-world adversarial behaviour. Additionally, graph construction step in the NIDS framework introduces a computational overhead that could be a constraint in high throughput, low-latency deployment scenarios. These

are directions where further work is needed.

5.5 Summary of Key Results

The two frameworks presented in this work tackle structurally different problems multimodal content moderation and network intrusion detection but reach similar conclusions about what makes deep learning effective in each domain.

For cyberbullying detection on the Facebook Hateful Meme Dataset, the proposed DR-HRNet + MC-DeBERTa + Adaptive EfficientNetV2 framework achieves 99.50% accuracy, with AUC values of 99.21% and 99.34% for the Bully and Non-Bully classes respectively. Confusion matrix shows only 4 total misclassifications across 1,000 test samples. All baseline models trail by at least 5 percentage points on accuracy, confirming that multimodal fusion is the key differentiator.

For network intrusion detection on CICIDS2017, the Hybrid GAT+BiLSTM+MLP model achieves 94.20% accuracy, 94.53% precision, and a macro-average AUC of 0.997 across five traffic classes. The confusion matrix concentrations along the diagonal confirm reliable per-class performance. The nearest competitor (GNN) is 3.55 percentage points lower on F1, and the weakest baseline (LSTM) is over 35 percentage points lower a gap large enough to make the architectural argument unambiguous.

Taken together, both results support the work that deep architectures tailored to the structure of the data multimodal for content, graph-plus-sequence for network traffic substantially outperform approaches that treat the problem as generic classification.

CHAPTER 6

CONCLUSION, FUTURE WORK, AND SOCIAL IMPACT

6.1 Conclusion

This work addressed two structurally different problems in cyber threat detection: identifying hateful multimodal content online, and classifying intrusion patterns in network traffic. The two have been studied separately for years, with tools developed for one rarely crossing over to the other. Here, both are treated as the same design question in different clothing how do you build a classifier that holds up when the data is noisy, imbalanced, and split across multiple information channels?

For cyberbullying, the approach was a multimodal pipeline: DR-HRNet for image features, MC-DeBERTa for text, Multimodal Compact Bilinear Pooling for cross-modal fusion, and Adaptive EfficientNetV2 with a Transformer Spatial Pooling module for classification. On the Facebook Hateful Meme Dataset, this reached **99.50%** accuracy on the Bully class and **99.20%** on Non-Bully, with AUC values of **0.9921** and **0.9934** respectively, and only 4 misclassifications across 1,000 test samples. Every baseline CNN-LSTM, BiGRU, CNet, VGG19, and MobileNetV2 trailed by at least 5 percentage points on accuracy.

For network intrusion detection, the approach was graph-based. Flows were modelled as nodes in a cosine-similarity kNN graph ($k = 10$, adaptive 70th-percentile threshold). Two stacked multi-head GAT layers learned topological attack patterns; a BiLSTM captured sequential flow evolution; an MLP performed five-class classification under a class-frequency-weighted Focal Loss. On CICIDS2017, this reached **94.20%** accuracy and F1-score with macro ROC-AUC of **0.997**, ahead of standalone GNN (90.65%), CNN (90.13%), MLP (72.58%), and LSTM (59.53%).

The common thread is that neither system works by using a larger model. The cyberbullying framework works because hateful memes are deliberately multimodal the offence often only registers when image and caption are read together, making MCB fusion the critical component. The NIDS framework works because network attacks

are relational and sequential flows do not occur in isolation and their order carries information. In both cases, fitting the architecture to the actual structure of the data, rather than applying a generic deep learning template, is what drives the performance margins.

6.2 Future Work

The limitation in the cyberbullying framework is modality coverage. The model handles image and text only, audio and video not included. Speech-based harassment, manipulated video memes and more becoming common on platforms such as TikTok. The framework is not tested on low resource language and code mixed cases. Languages especially Romanized Bangla, Urdu or Thai that have informal styles that differ from those of English-language social media. Any Such system used for world-wide deployment should be capable of coping with this.

For the NIDS framework issue is real-time adaptability. The model is based on a static graph that is constructed for a set of features matrix. If deployed live in a network operations centre then would require dynamic graph construction. There is also need of sliding window flow buffering and incremental updates to kNN. Edge deployment is an issue as well. The devices have an inference running. To comply with this requirement, the network switches are kept near the network, rather than in the central data centre. It's not designed to do this.

The frameworks might be able to be improved by incorporating Explainable AI. High accuracy Securing and identifying "content" is not sufficient for security and content moderation. Analysts need to Be able to explain the rationale for a classification. In the case of cyberbullying it means knowing which regions of the image caused the prediction. For NIDS, it is knowing which flow nodes and graph edges caused the alert, and for the image. SHAP, LIME, attention rollout, and GNN explainability methods are the most promising directions. Without interpretable outputs, even a highly accurate system will face adoption barriers from reviewers who are reluctant, reasonably, to act on unexplained flags.

Federated learning is worth pursuing in both domains. Centralising sensitive network traffic or user-generated content for training is legally and ethically constrained; federated approaches would allow collaborative improvement without pooling the data itself. And both frameworks need continual learning mechanisms the ability to adapt to new meme formats, emerging attack types, and shifting traffic patterns without full

retraining. Cyber threats change over time, and a system that cannot update will degrade. That is probably the most important open problem this work leaves behind.

6.3 Social Impact

The numbers in this work correspond to real problems. Cyberbullying causes measurable psychological harm depression, anxiety, academic failure, and in documented cases, suicidal ideation hitting children, adolescents, and marginalised communities hardest. The FHM Dataset exists because text-only and image-only filters kept failing on this content category. A system that reaches 99.50% accuracy on Bully content and AUC above 0.99 is close enough to deployment. Its use in a content moderation pipeline becomes a real possibility. On major platforms millions of posts are published every hour, even small improvement in detection rate removes a large volume of harmful content before it reaches users particularly younger users, who are statistically most at risk.

The NIDS framework focus on different types of threat. DDoS and DoS attacks take essential services offline; PortScan is typically the early step before something worse. A classifier that handles five traffic categories and achieves 94.20% F1 and 0.997 macro AUC. It maintains minority-class recall strong through dual imbalance mitigation, gives security analysts a better alternative than single-paradigm alternatives. It operates on flow-level features without payload inspection, making it compatible with privacy-preserving network architectures . The data protection regulations that increasingly restrict deep packet inspection.

Both systems reflect the same underlying point: the useful safety tools are ones that work under conditions that actually exist. Real social media data is noisy, multilingual, and multimodal. Real network traffic is imbalanced and encrypted. Both frameworks were built with those realities in mind.

REFERENCES

[1] A. Fitro, M. A. Wibowo, and C. E. Widodo, "Automatic detection of cyberbullying on text, image, and video: A systematic literature review," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 15, no. 01, pp. 75–80, 2026.

[2] L. Diana, P. Dini, and D. Paolini, "Overview on intrusion detection systems for computer networking security," *Computers*, vol. 14, p. 87, Mar. 2025.

[3] C. Kumari and M. Kaur, "XMTF: Cross-modal fusion with DINOv2 for cyberbullying detection in code-mixed memes," in *International Conference on Data Analytics & Management*, pp. 436–445, Springer Nature Switzerland, 2025.

[4] S. Tahir and A. Nawaz, "HUEID: Hybrid deep learning for cyberbullying detection using multi-modal Urdu text and emojis," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 11, pp. 1–19, 2025.

[5] T. Mahmud and A. C. Roy, "A multimodal multiclass cyberbullying classification in romanized bangla social media content," *Proceedings of the 28th International Conference on Computer and Information Technology (ICCIT)*, Dec. 2025.

[6] A. G. Philipo, D. S. Sarwatt, J. Ding, M. Daneshmand, and H. Ning, "Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms," *ACM Computing Surveys*, 2024.

[7] S. Sattar, S. Khan, M. I. Khan, A. Akhmediyarova, O. Mamyrbayev, D. Kassymova, D. Oralbekova, and J. Alimkulova, "Anomaly detection in encrypted network traffic using self-supervised learning," *Scientific Reports*, vol. 15, p. 26585, July 2025.

[8] E. Papadogiannaki and S. Ioannidis, "A survey on encrypted network traffic analysis applications, techniques, and countermeasures," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[9] S. Mohammedjany *et al.*, "Detecting multimodal cyber-bullying behaviour in social-media using deep learning techniques," *The Journal of Supercomputing*, vol. 81, no. 1, p. 284, 2025.

[10] N. M. Singh and S. K. Sharma, "An efficient automated multi-modal cyberbullying detection using decision fusion classifier on social media platforms," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 20507–20535, 2024.

[11] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21954–21961, 2017.

[12] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1690–1700, 2014.

[13] P. Jha, K. Maity, R. Jain, A. Verma, S. Saha, and P. Bhattacharyya, "Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, (St. Julian's, Malta), pp. 930–943, Association for Computational Linguistics, Mar. 2024.

[14] G. Yang *et al.*, "Cross-lingual and multimodal cyberbullying and bias detection and content generation via CyberGenDet," in *Intelligent Systems Conference*, pp. 210–224, Springer Nature Switzerland, 2025.

[15] M. S. Karthikeyan *et al.*, "Towards safer digital spaces: Automated detection of cyberbullying through multi-modal learning," in *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, pp. 480–486, IEEE, 2024.

[16] F. B. F. Ananna *et al.*, "Cyberbullying detection: An innovative multitask learning framework for analyzing harmful memes," in *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, pp. 1–6, IEEE, 2024.

[17] K. Luo, C. Zheng, and Z. Guan, "Reinforced multi-modal cyberbullying detection with subgraph neural networks," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 3, pp. 2161–2180, 2025.

[18] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets, and challenges," *Cybersecurity*, vol. 2, no. 1, p. 20, 2019.

[19] S. T. Ikram, A. K. Cherukuri, B. Poorva, P. S. Ushasree, Y. Zhang, X. Liu, and G. Li, "Anomaly detection using XGBoost ensemble of deep neural network models," *Cybernetics and Information Technologies*, vol. 21, pp. 175–188, Sept. 2021.

[20] F. J. ALzaher and A. AlJarullah, "Intrusion detection using machine learning and deep learning," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 8, 2025.

[21] H. R. Ibraheem, N. D. Zaki, and M. I. Al-Mashhadani, "Anomaly detection in encrypted HTTPS traffic using machine learning: A comparative analysis of feature selection techniques," *Mesopotamian Journal of Computer Science*, pp. 18–28, 2022.

[22] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *Journal of Information Security and Applications*, vol. 50, p. 102419, 2020.

[23] M.-G. Kim and H. Kim, "Anomaly detection in imbalanced encrypted traffic with few packet metadata-based feature extraction," *CMES - Computer Modeling in Engineering and Sciences*, vol. 141, no. 1, pp. 585–607, 2024.

[24] P. Goldschmidt and D. Chudá, "Network intrusion datasets: A survey, limitations, and recommendations," *Computers & Security*, p. 104510, 2025.

[25] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[26] N. Sultana, N. Chilamkurti, W. Peng, and R. Alhadad, "Survey on SDN-based network intrusion detection systems using machine learning approaches," *Peer-to-Peer Networking and Applications*, vol. 12, no. 2, pp. 493–501, 2019.

[27] H. Khafajeh, "Cyberbullying detection in social networks using deep learning," *International Arab Journal of Information Technology*, vol. 21, no. 6, 2024.

[28] M. A. Arshed *et al.*, "Multi-class visual cyberbullying detection using deep neural networks and the CVID dataset," *Information*, vol. 16, no. 8, p. 630, 2025.

[29] I. Tabassum and V. Nunavath, "A deep-learning based approach for multi-class cyberbullying classification using social media text and image data," in *Norsk IKT-konferanse for forskning og utdanning*, no. 2, 2024.

[30] T. Jaiswal and V. K. Attri, "A proposed multi-modal approach to detect cyberbullying in multi file format – an empirical study," in *International Conference on Smart Systems and Wireless Communication*, pp. 541–557, Springer Nature Singapore, 2024.

[31] I. Musyoka, J. Wandeto, and B. Kituku, "Integrating BERT and RESNET50V2 for multimodal cyberbullying detection*," in *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*, pp. 1–6, 2024.

[32] M. A. Al-Khasawneh, M. Faheem, A. A. Alarood, S. Habibullah, and E. Alsolami, "Toward multi-modal approach for identification and detection of cyberbullying in social networks," *IEEE Access*, vol. 12, pp. 90158–90170, 2024.

[33] S. Phosit, S. Kongsamlit, and K. Pasupa, "Detecting cyberbullying in Thai memes: A multimodal approach using deep learning," in *2025 IEEE Symposium on Computational Intelligence in Natural Language Processing and Social Media (CINLP-SoMe)*, pp. 1–7, IEEE, 2025.

[34] N. M. Singh and S. K. Sharma, "Multi-modal cyberbullying detection with severity analysis using deep-tensor fusion framework," *International Journal of Computer Network and Information Security*, vol. 3, pp. 144–150, 2025.

[35] T. Li, Z. Zeng, Q. Li, and S. Sun, "Integrating GIN-based multimodal feature transformation and multi-feature combination voting for irony-aware cyberbullying detection," *Information Processing & Management*, vol. 61, no. 3, p. 103651, 2024.

[36] A. C. Roy, T. Mahmud, and T. Abrar, "A multi-class cyberbullying classification on image and text in code-mixed bangla-english social media content," *Natural Language Processing Journal*, p. 100191, 2025.

[37] K. Singh, A. Kashyap, and A. K. Cherukuri, "Interpretable anomaly detection in encrypted traffic using SHAP with machine learning models," *arXiv preprint arXiv:2505.16261*, May 2025.

[38] M. Mozaffari, K. Doshi, and Y. Yilmaz, "Self-supervised learning for online anomaly detection in high-dimensional data streams," *Electronics*, vol. 12, no. 9, p. 1971, 2023.

[39] M. Sajid *et al.*, "Enhancing intrusion detection: A hybrid machine and deep learning approach," *Journal of Cloud Computing*, vol. 13, no. 1, p. 123, 2024.

[40] M. A. Alsoufi *et al.*, "Anomaly-based intrusion detection model using deep learning for IoT networks," *Computer Modeling in Engineering & Sciences*, vol. 141, no. 1, pp. 823–845, 2024.

[41] C. Fu, Q. Li, and K. Xu, "Detecting unknown encrypted malicious traffic in real time via flow interaction graph analysis," *arXiv preprint arXiv:2301.13686*, Jan. 2023.

[42] A. Ferriyan, A. H. Thamrin, K. Takeda, and J. Murai, "Generating network intrusion detection datasets based on real and encrypted synthetic attack traffic," *Applied Sciences*, vol. 11, no. 17, p. 7868, 2021.

[43] R. T. Elmaghraby, N. M. A. Aziem, M. A. Sobh, and A. M. Bahaa-Eldin, "Encrypted network traffic classification based on machine learning," *Ain Shams Engineering Journal*, vol. 15, no. 2, p. 102361, 2024.

[44] Canadian Institute for Cybersecurity, "CICIDS2017: Intrusion detection evaluation dataset." `https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset`, 2017.

[45] A. N. Gummadi, J. C. Napier, and M. Abdallah, "XAI-IoT: An explainable AI framework for enhancing anomaly detection in IoT systems," vol. 12, pp. 71024–71054, 2024.

[46] M. D. Nguyen, V. H. La, W. Mallouli, A. R. Cavalli, and E. M. de Oca, "Toward anomaly detection using explainable AI," in *CyberSecurity in a DevOps Environment: From Requirements to Monitoring*, pp. 293–324, Springer, 2023.

[47] Kaggle, "Facebook hateful meme dataset." `https://www.kaggle.com/datasets/parthplc/facebook-hateful-meme-dataset`, 2024.

[48] B. Sun, Q. Li, and Z. Zhang, "Dynamic context modeling based lightweight high-resolution network for dense prediction," *Engineering Applications of Artificial Intelligence*, vol. 129, p. 107642, 2024.

[49] F. Li, S. Guo, F. Han, J. Zhao, and F. Shen, "Multi-scale dilated convolution network for long-term time series forecasting," *arXiv preprint arXiv:2405.05499*, May 2024.

[50] M. M. Ahsan, T. E. Alam, T. Trafalis, and P. Huebner, "Deep MLP-CNN model using mixed-data to distinguish between COVID-19 and non-COVID-19 patients," *Symmetry*, vol. 12, p. 1526, Sept. 2020.

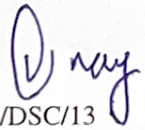
[51] I. Pacal *et al.*, "Enhancing EfficientNetV2 with global and efficient channel attention mechanisms for accurate MRI-based brain tumor classification," *Cluster Computing*, vol. 27, no. 8, pp. 11187–11212, 2024.

DECLARATION

We hereby certify that the work which is presented in the Major Project-II/Research Work entitled **"Hybrid Deep Learning Approaches for Cyber Threat Detection in Network and Social Media"** in fulfillment of the requirement for the award of the Degree of Master of Technology in **Data Science** and submitted to the Department of **Software Engineering**, Delhi Technological University, Delhi, is an authentic record of our own work, carried out during a period from **August 2024 - May 2026** under the supervision of **Dr. Rahul**.

The matter presented in this report/thesis has not been submitted by us for the award of any other degree of this or any other Institute/University. The work has been published/accepted/communicated in SCI/SCI-expanded/SSCI/Scopus indexed journal OR peer-reviewed Scopus indexed conference with the following details:

Title of the Paper: Hybrid Multimodal Image-Text Classification for Cyberbullying Detection using DR-HRNet and MC-DeBERTa with Adaptive EfficientNetV2
Author Names (in sequence as per research paper): Vinay Rana, Dr. Rahul
Name of Conference/Journal: 7th International Conference on Data Analytics & Management (ICDAM-2026)
Conference Dates with venue (if applicable): 12-14 June 2026, London Metropolitan University, London, UK
Have you registered for the conference?: Yes
Status of paper (Accepted/Published/Communicated): Accepted (Scopus-Indexed)
Date of paper Communication: 29 April 2026
Date of paper Acceptance: 6 May 2026
Date of paper Publication: To be published



24/DSC/13
VINAY RANA
Student(s) Roll No., Name and Signature

SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I further certify that the publication and indexing information given by the student is correct.

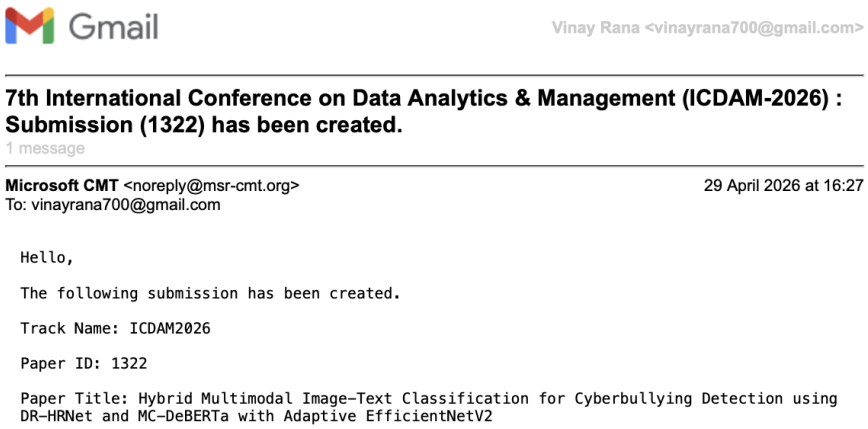
Place: Delhi
Date: May 21, 2026

Dr. Rahul
Supervisor Name and Signature

NOTE: PLEASE ENCLOSE RESEARCH PAPER ACCEPTANCE/PUBLICATION/COMMUNICATION PROOF ALONG WITH SCOPUS INDEXING PROOF (conference website or ScienceDirect in case of journal publication).

Rana, V. and Rahul, *"Hybrid Multimodal Image-Text Classification for Cyberbullying Detection using DR-HRNet and MC-DeBERTa with Adaptive EfficientNetV2"*. The paper has been **Accepted** at the **7th International Conference on Data Analytics & Management (ICDAM-2026)**, to be held at **London Metropolitan University , London, UK**, during **12-14 June 2026**. The conference proceedings are indexed by **Scopus**. Paper ID: **1322**.

Submission:



Gmail Vinay Rana <vinayrana700@gmail.com>

7th International Conference on Data Analytics & Management (ICDAM-2026) : Submission (1322) has been created.

1 message

Microsoft CMT <noreply@msr-cmt.org> 29 April 2026 at 16:27
To: vinayrana700@gmail.com

Hello,

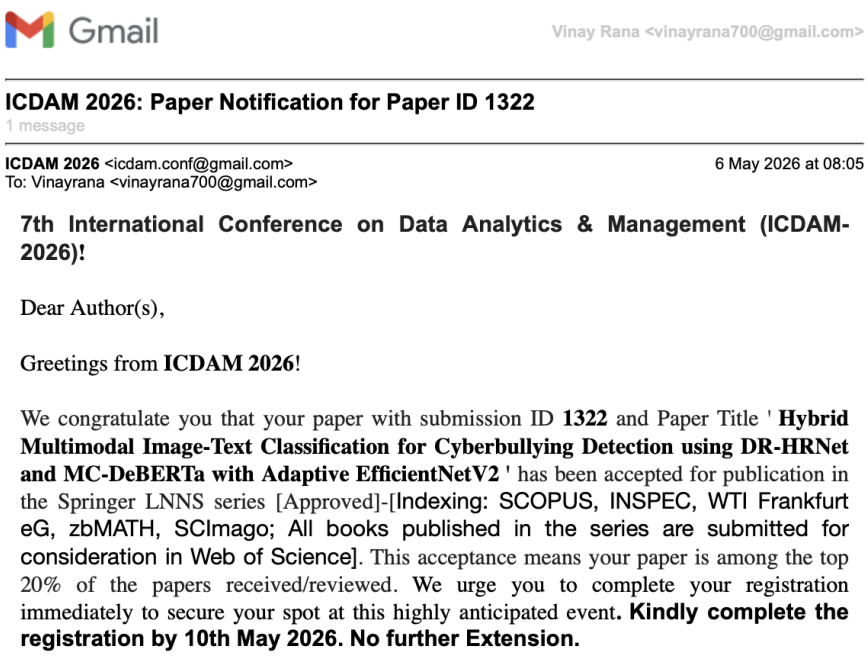
The following submission has been created.

Track Name: ICDAM2026

Paper ID: 1322

Paper Title: Hybrid Multimodal Image-Text Classification for Cyberbullying Detection using DR-HRNet and MC-DeBERTa with Adaptive EfficientNetV2

Acceptance:



Gmail Vinay Rana <vinayrana700@gmail.com>

ICDAM 2026: Paper Notification for Paper ID 1322

1 message

ICDAM 2026 <icdam.conf@gmail.com> 6 May 2026 at 08:05
To: Vinayrana <vinayrana700@gmail.com>

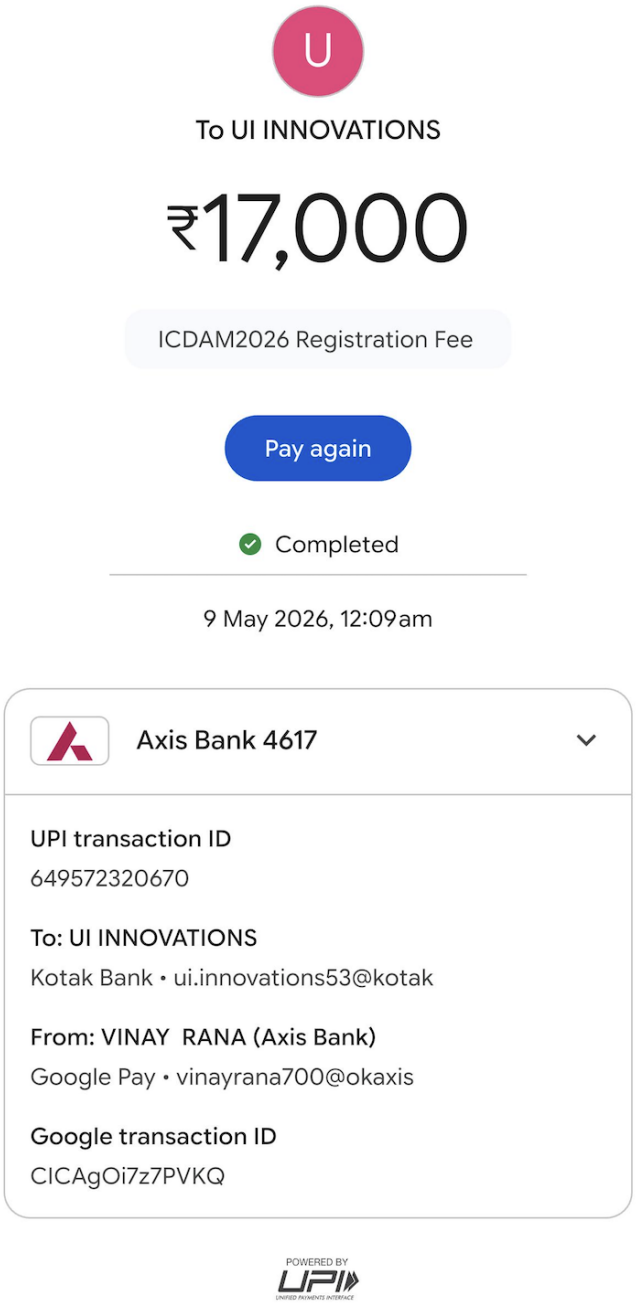
7th International Conference on Data Analytics & Management (ICDAM-2026)!

Dear Author(s),

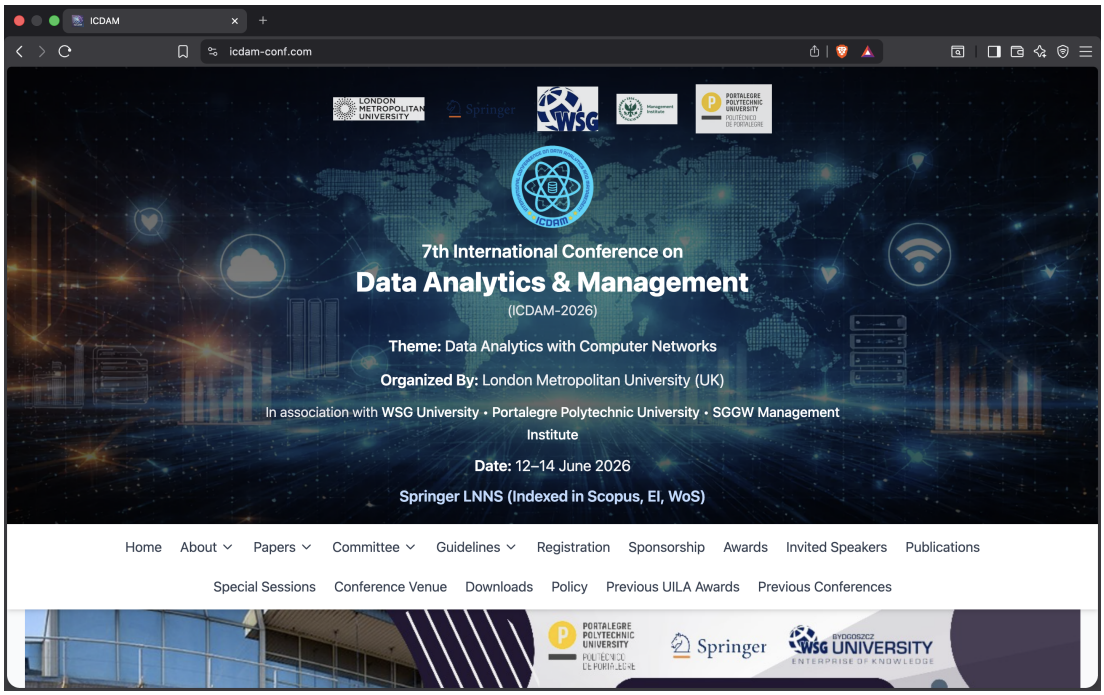
Greetings from **ICDAM 2026**!

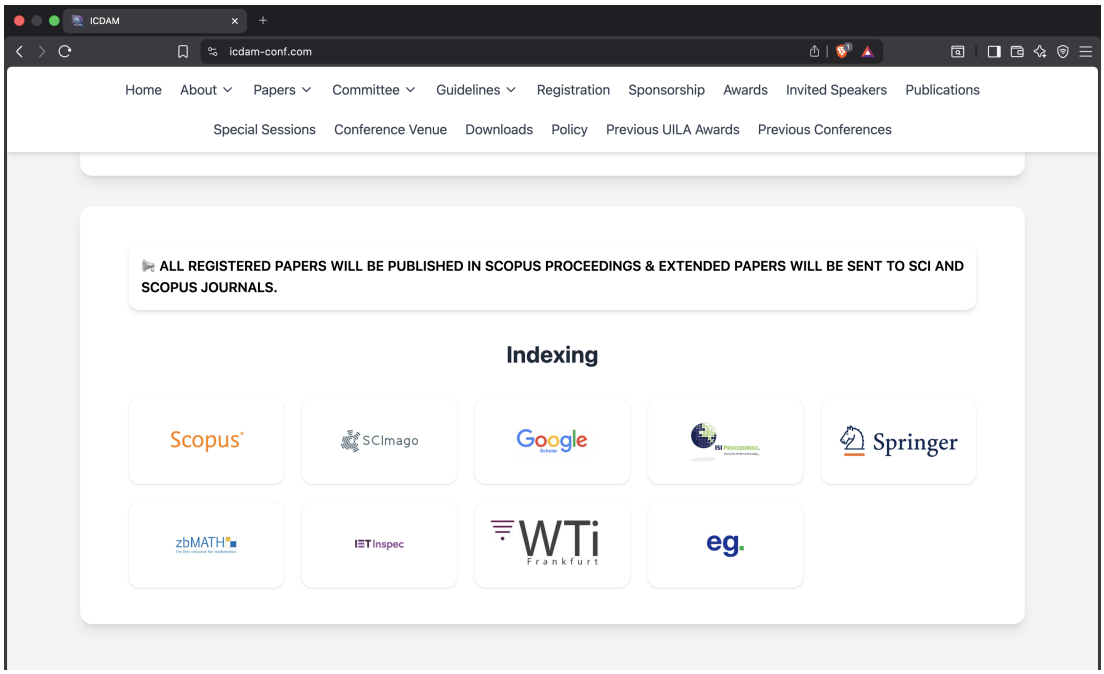
We congratulate you that your paper with submission ID **1322** and Paper Title ' **Hybrid Multimodal Image-Text Classification for Cyberbullying Detection using DR-HRNet and MC-DeBERTa with Adaptive EfficientNetV2** ' has been accepted for publication in the Springer LNNS series [Approved]-[Indexing: SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago; All books published in the series are submitted for consideration in Web of Science]. This acceptance means your paper is among the top 20% of the papers received/reviewed. We urge you to complete your registration immediately to secure your spot at this highly anticipated event**. Kindly complete the registration by 10th May 2026. No further Extension.**

Payment Proof:



Scopus Proof:



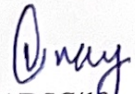


DECLARATION

We hereby certify that the work which is presented in the Major Project-II/Research Work entitled **"Hybrid Deep Learning Approaches for Cyber Threat Detection in Network and Social Media"** in fulfillment of the requirement for the award of the Degree of Master of Technology in **Data Science** and submitted to the Department of **Software Engineering**, Delhi Technological University, Delhi, is an authentic record of our own work, carried out during a period from **August 2024 - May 2026** under the supervision of **Dr. Rahul**.

The matter presented in this report/thesis has not been submitted by us for the award of any other degree of this or any other Institute/University. The work has been published/accepted/communicated in SCI/SCI-expanded/SSCI/Scopus indexed journal OR peer-reviewed Scopus indexed conference with the following details:

Title of the Paper: Graph-Based Hybrid GAT-BiLSTM Framework For Robust Intrusion Detection In Imbalanced Network Traffic
Author Names (in sequence as per research paper): Vinay Rana, Dr. Rahul
Name of Conference/Journal: 7th International Conference on Data Analytics & Management (ICDAM-2026)
Conference Dates with venue (if applicable): 12-14 June 2026, London Metropolitan University, London, UK
Have you registered for the conference?: Yes
Status of paper (Accepted/Published/Communicated): Accepted (Scopus-Indexed)
Date of paper Communication: 29 April 2026
Date of paper Acceptance: 6 May 2026
Date of paper Publication: To be published



24/DSC/13
VINAY RANA
Student(s) Roll No., Name and Signature

SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I further certify that the publication and indexing information given by the student is correct.

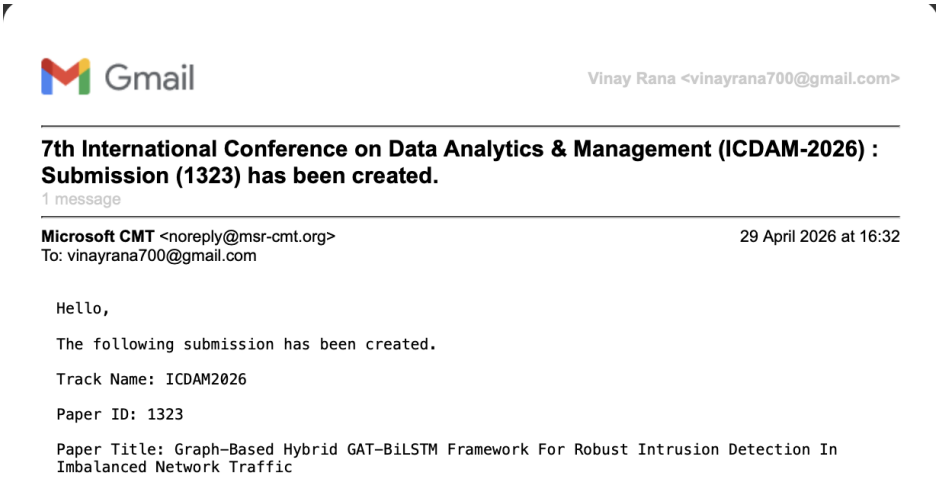
Place: Delhi
Date: May 21, 2026

Dr. Rahul
Supervisor Name and Signature

NOTE: PLEASE ENCLOSE RESEARCH PAPER ACCEPTANCE/PUBLICATION/COMMUNICATION PROOF ALONG WITH SCOPUS INDEXING PROOF (conference website or ScienceDirect in case of journal publication).

Rana, V. and Rahul, *"Graph-Based Hybrid GAT-BiLSTM Framework For Robust Intrusion Detection In Imbalanced Network Traffic"*. The paper has been **Accepted** at the **7th International Conference on Data Analytics & Management (ICDAM-2026)**, to be held at **London Metropolitan University , London, UK**, during **12-14 June 2026**. The conference proceedings are indexed by **Scopus**. Paper ID: **1323**.

Submission:



Gmail Vinay Rana <vinayrana700@gmail.com>

7th International Conference on Data Analytics & Management (ICDAM-2026) : Submission (1323) has been created.

1 message

Microsoft CMT <noreply@msr-cmt.org> 29 April 2026 at 16:32
To: vinayrana700@gmail.com

Hello,

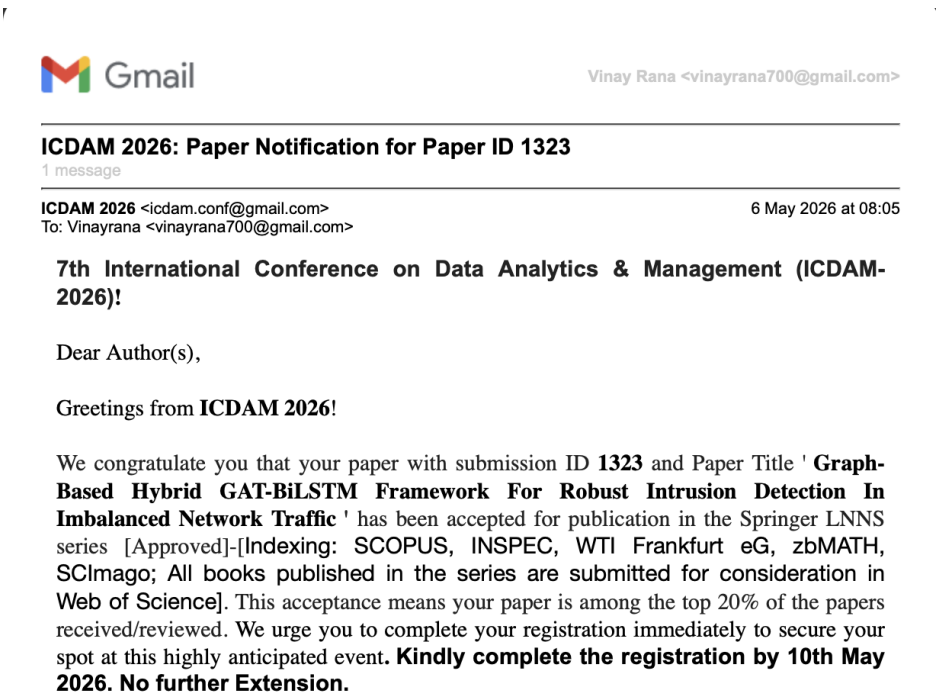
The following submission has been created.

Track Name: ICDAM2026

Paper ID: 1323

Paper Title: Graph-Based Hybrid GAT-BiLSTM Framework For Robust Intrusion Detection In Imbalanced Network Traffic

Acceptance:



Gmail Vinay Rana <vinayrana700@gmail.com>

ICDAM 2026: Paper Notification for Paper ID 1323

1 message

ICDAM 2026 <icdam.conf@gmail.com> 6 May 2026 at 08:05
To: Vinayrana <vinayrana700@gmail.com>

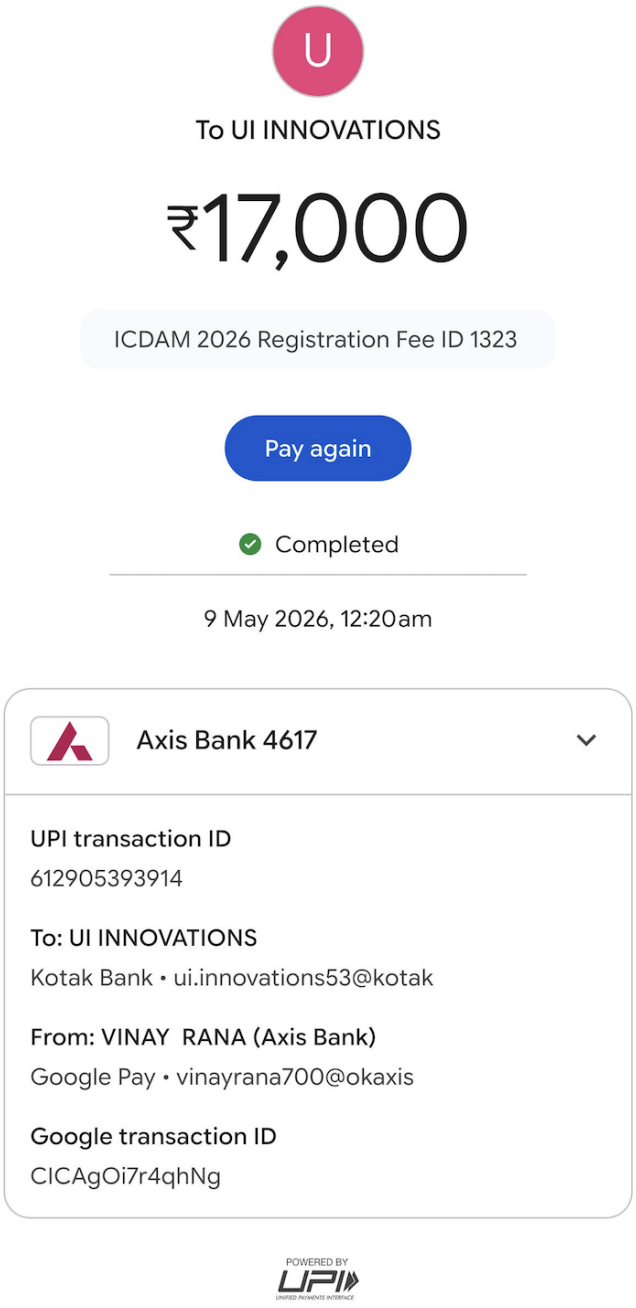
7th International Conference on Data Analytics & Management (ICDAM-2026)!

Dear Author(s),

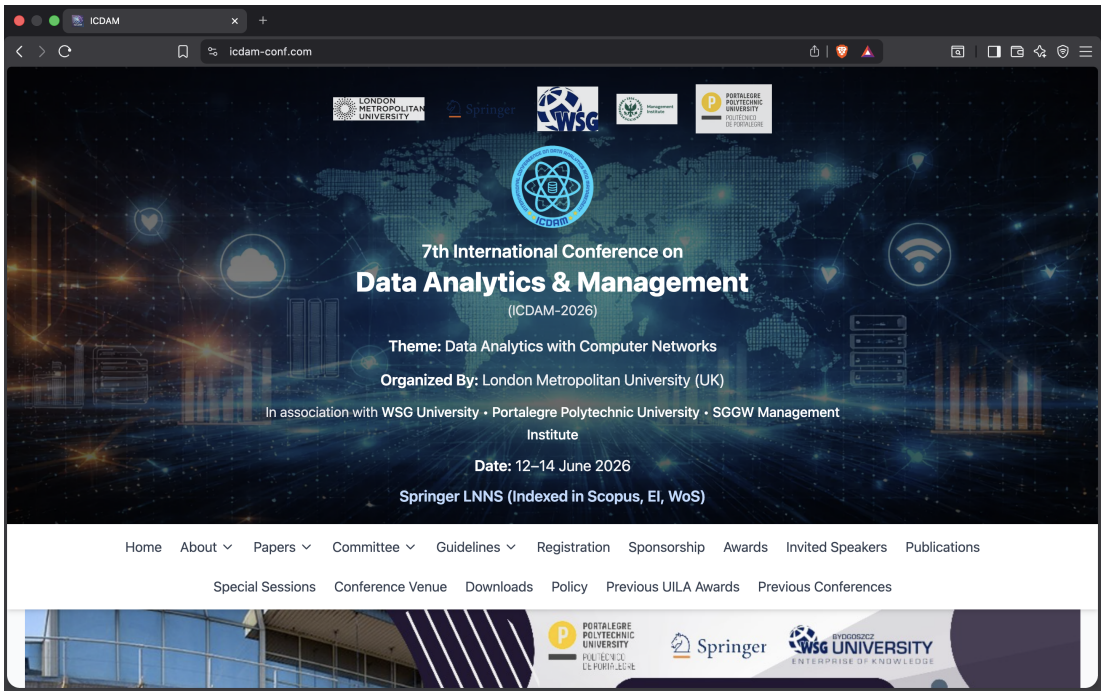
Greetings from **ICDAM 2026**!

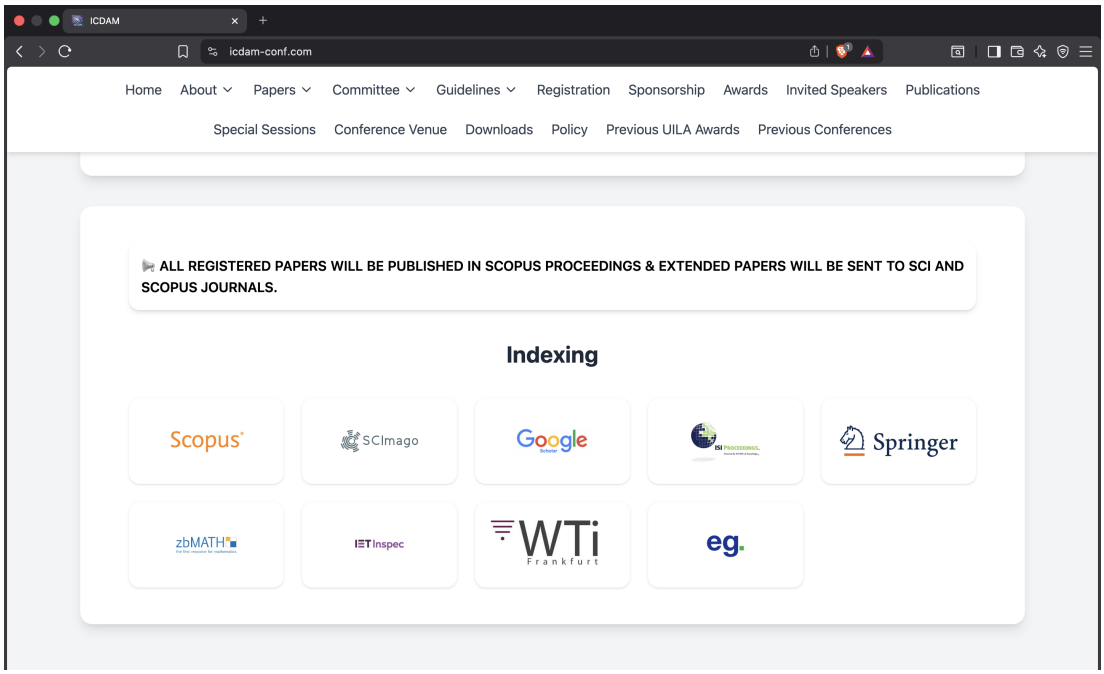
We congratulate you that your paper with submission ID **1323** and Paper Title ' **Graph-Based Hybrid GAT-BiLSTM Framework For Robust Intrusion Detection In Imbalanced Network Traffic** ' has been accepted for publication in the Springer LNNS series [Approved]-[Indexing: SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago; All books published in the series are submitted for consideration in Web of Science]. This acceptance means your paper is among the top 20% of the papers received/reviewed. We urge you to complete your registration immediately to secure your spot at this highly anticipated event**. Kindly complete the registration by 10th May 2026. No further Extension.**

Payment Proof:



Scopus Proof:





DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis HYBRID DEEP LEARNING APPROACHES FOR CYBER THREAT DETECTION IN NETWORK AND SOCIAL MEDIA

Total Pages 77 Name of the Scholar VINAY RANA

Supervisor(s)

(1) Dr. RAHUL

(2) -

(3) -

Department SOFTWARE ENGINEERING

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: TURNITIN Similarity Index: 6% , Total Word Count: 18,126

Date: 29/5/26

Candidate's Signature **Signature of Supervisor(s)**

Vinay Rana

VINAY_RANA_THESIS_24_DSC_13

M.TECH MAJOR PROJECT II

Document Details

Submission ID
trn:oid:::27535:140805541

Submission Date
May 28, 2026, 6:33 PM GMT+5:30

Download Date
May 28, 2026, 6:39 PM GMT+5:30

File Name
VINAY_RANA_THESIS_24_DSC_13.pdf

File Size
11.4 MB

77 Pages

18,126 Words

104,870 Characters

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 10 words)

Match Groups

55 Not Cited or Quoted 5%
Matches with neither in-text citation nor quotation marks

10 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
2% Publications
5% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **55** Not Cited or Quoted 5%
  Matches with neither in-text citation nor quotation marks
- **10** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
2% Publications
5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1 Student papers
University of Edinburgh on 2011-08-17 <1%

2 Student papers
Brunel University on 2025-07-23 <1%

3 Student papers
Addis Ababa University on 2025-06-23 <1%

4 Internet
www.mdpi.com <1%

5 Student papers
University of Hong Kong on 2020-05-06 <1%

6 Student papers
VIT University on 2025-04-11 <1%

7 Internet
ir.uitm.edu.my <1%

8 Student papers
George Mason University on 2026-05-25 <1%

9 Student papers
University of Reading on 2024-10-11 <1%

10 Internet
ebin.pub <1%

11 Internet
raw.githubusercontent.com <1%

12 Internet
www.voltcoffer.com <1%

13 Student papers
Strathmore University (Main Account) on 2026-03-24 <1%

14 Student papers
University of East London on 2023-05-10 <1%

15 Internet
fdocumenti.com <1%

16 Publication
Anwar Hossain Efat. "Tri-Attention boosted scalable EfficientNet for skin lesion cl... <1%

17 Student papers
Teesside University (Blackboard Ultra) on 2026-05-01 <1%

18 Internet
ijettjournal.org <1%

19 Publication
Jiayu Wang, Xuehu Yan, Lintao Liu, Longlong Li, Yongqiang Yu. "CTTGAN: Traffic D... <1%

20 Student papers
UNICAF on 2025-04-28 <1%

21 Publication
"Proceedings of the 6th International Conference on Artificial Intelligence and S... <1%

22 Student papers
Sheffield Hallam University on 2026-01-07 <1%

23 Internet
fondazione.cogeme.net <1%

24 Internet
www.research.unipd.it <1%

25 Publication
"Computer Vision – ECCV 2016", Springer Nature, 2016 <1%

26 Student papers
Brunel University on 2026-04-27 <1%

27 Publication
Shouqiang Sun, Ziming Zeng, Qingqing Li, Tingting Li. "A fine-grained scene retrie... <1%

28 Publication
Shuning Ge, Fangyun Qin, Xiaohui Wan, Yang Liu, Qian Dai, Zheng Zheng. "ARFT-T... <1%

29 Student papers
University of Bolton on 2026-05-17 <1%

30 Student papers
University of Wollongong on 2018-10-01 <1%

31 Internet
d197for5662m48.cloudfront.net <1%

32 Internet
docslib.org <1%

33 Internet
escholarship.org <1%

34 Internet
tus.repo.nii.ac.jp <1%

35 Internet
unige.iris.cineca.it <1%

36 Publication
Angel A. Rodriguez, Mason Davis, Joshua Zander, Edwin Nazario Dejesus, Moham... <1%

37 Student papers
Carnegie Mellon University on 2026-02-23 <1%

38 Publication
Chen, Wenyuan. "AI-Based Subcellular Morphology Measurement for Sperm.", Un... <1%

39 Student papers
Engineering Faculty on 2026-05-04 <1%

40 Student papers
Liverpool John Moores University on 2026-05-12 <1%

41 Student papers
Teesside University (Blackboard Ultra) on 2026-05-01 <1%

42 Student papers
University of Ulster on 2026-04-19 <1%

43 Student papers
University of West London on 2025-06-14 <1%

44 Internet
arxiv.org <1%

45 Internet
psc.ky.gov <1%

46 Internet
www.preprints.org <1%

47 Publication
"Proceedings of 5th International Conference on Frontiers in Computing and Syst... <1%

48 Publication
Ahmed A. Abd El-Latif, Mohammed A ElAffendi, Mohamed Ali AlShara, Yassine Ma... <1%

49 Student papers
Ahsanullah University of Science and Technology on 2026-04-06 <1%

50 Student papers
Dayananda Sagar College of Engg., Bengaluru on 2026-05-06 <1%

51 Student papers
Information Technology on 2026-05-11 <1%

52 Publication
Lecture Notes in Electrical Engineering, 2015. <1%

53 Publication

Md. Abdur Rahim, Md Mustafizur Rahman, Md Shofiqul Islam, Chukwudi Isinka Jo... <1%

54 Student papers

Moodle2025 on 2026-05-02 <1%

55 Student papers

Sharda University on 2026-03-30 <1%

56 Publication

Souhail Bakkali, Zuheng Ming, Mickaël Coustaty, Marçal Rusiñol. "EAML: ensembl... <1%

57 Student papers

Teesside University (Blackboard Ultra) on 2026-01-06 <1%

58 Student papers

University of Engineering & Technology Mardan on 2025-09-22 <1%

59 Internet

d-nb.info <1%

Vinay Rana

VINAY_RANA_THESIS_24_DSC_13

M.TECH MAJOR PROJECT II

Document Details

Submission ID
trn:oid:::27535:140805541

Submission Date
May 28, 2026, 6:33 PM GMT+5:30

Download Date
May 28, 2026, 6:39 PM GMT+5:30

File Name
VINAY_RANA_THESIS_24_DSC_13.pdf

File Size
11.4 MB

77 Pages

18,126 Words

104,870 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution! Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

1 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

AUTHOR'S PROFILE



Vinay Rana is currently pursuing the Master of Technology (M.Tech.) degree in Data Science from the Department of Software Engineering, Delhi Technological University (DTU), Delhi, India. His research work focuses on deep learning approaches for cyber threat detection, network intrusion detection systems, multimodal cyberbullying detection, and intelligent cybersecurity frameworks. His thesis entitled *"Hybrid Deep Learning Approaches for Cyber Threat Detection in Network and Social Media"* investigates graph-based intrusion detection models, multimodal cyberbullying detection frameworks, transformer-based architectures, and hybrid deep learning techniques for robust cybersecurity applications.

His research interests include Cybersecurity, Deep Learning, Network Intrusion Detection Systems (NIDS), Cyberbullying Detection, Graph Neural Networks, Transformer-based Architectures, Multimodal Learning, Explainable Artificial Intelligence, and Intelligent Threat Analytics. During his postgraduate research, he has worked on graph-based intrusion detection using hybrid GAT-BiLSTM architectures and multimodal social media threat detection using image-text fusion frameworks integrating transformer and convolutional neural network models.

He has contributed to research publications that have been accepted in international conferences in the domains of cybersecurity, network intrusion detection, cyberbullying detection, and artificial intelligence. His research contributions include the development of a Graph-Based Hybrid GAT-BiLSTM framework for robust intrusion detection in imbalanced network traffic and a Hybrid Multimodal Image-Text Classification framework for cyberbullying detection using DR-HRNet and MC-DeBERTa architectures.