Exploring Supervised Machine Learning Models (LSTM) on Multimodal Data to Detect Distracted Students in Immersive Educational Virtual Reality Environment

Subigya Gautam

A Thesis Presented to the Graduate Faculty
In Partial Fulfillment of the Requirements for the Degree
Master of Science

University of Louisiana at Lafayette
Spring 2025

APPROVED:

Arun K. Kulshresthh, Chair
School of Computing and Informatics

Christoph W. Borst
School of Computing and Informatics

Michael W. Totaro
School of Computing and Informatics

Mary Farmer-Kaiser
Dean of the Graduate School

Abstract

Distractions in VR education are a shift in cognitive focus that disrupts user engagement, it can be caused by either external stimuli (e.g., noise in the environment) or internal psychological state (e.g., mind wandering). Since distractions are often not apparent in VR, it is hard to identify them. Current VR systems lack efficient techniques to detect these cognitive shifts, limiting their ability to respond to changing user engagement.

This study explores three modeling approaches using an LSTM-based model that integrates EEG, VR, and eye gaze data to classify distractions. The approaches include generalized, participant-independent, and personalized models. The generalized model performed well (F1: 0.95, accuracy: 0.93), participant-independent models showed high variability, 63% scored F1 < 0.6 with only 11.1% exceeding 0.9. Personalized models showed dramatic performance differences, some were near-perfect (F1 > 0.95) while others failed (F1 = 0). The distraction patterns were highly individualized. This variability shows the need for a user-specific optimization approach that accommodates individual distraction patterns.

Despite challenges, personalized models demonstrated strong potential, indicating that adaptive ML can improve distraction detection. Future research should explore hybrid methods that combine generalized baselines with personalized calibration to enhance reliability. Addressing individual differences in cognitive engagement leads to more effective VR-based learning solutions.

Whatever happened, happened for the good. Whatever is happening, is happening for the good. Whatever will happen, will also happen for the good.

“If you optimize everything, you will always be unhappy. And the worst flaw is that we’re just plain dumb. Admit it!” — Donald Knuth

Acknowledgments

With profound gratitude and heartfelt appreciation, I sincerely thank Dr. Arun Kulshreshth for his unwavering guidance, invaluable insight, and steadfast support throughout this journey. His mentorship has been the cornerstone of my academic growth, and much of my success is due to his wisdom and encouragement.

I am equally grateful to my thesis committee members, Dr. Christoph Borst and Dr. Michael Totaro, for their thoughtful reviews and insightful feedback that have profoundly shaped this work. My sincere thanks also go to the incredible members of the HCI Lab, whose collaborative spirit and generosity have made this journey both enriching and fulfilling. A special mention goes to Nicholas Fisher for his support in setting up the research environment.

Beyond academia, my heartfelt appreciation extends to my friends, whose direct and indirect support has strengthened and motivated me. And above all, I am forever grateful to my family. Mom and Dad, your sacrifices, love, and unwavering belief in me have made this achievement possible. Every milestone I reach is a testament to your endless support, and I owe everything to you.

This journey has been a collective effort, shaped by the kindness, wisdom, and encouragement of so many. To everyone who has been part of it, I am deeply grateful.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
4D	Four-Dimensional
AI	Artificial Intelligence
AR	Augmented Reality
AUC	Area Under the Curve
BCI	Brain-Computer Interface
BPM	Beats Per Minute
CA	Content Anomaly
CSV	Comma-Separated Values
CV	Computer Vision
EEG	Electroencephalograph
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HCI	Human-Computer Interaction
HMD	Head Mounted Device
HR	Heart Rate
Hz	Hertz
ID	Identifier

IRB	Institutional Review Board
LLM	Large Language Model
LSTM	Long Short-Term Memory
MM	Millimeter
MR	Mixed Reality
PC	Personal Computer
PR	Precision-Recall
RAM	Random Access Memory
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RTX	Ray Tracing Texel eXtreme
SD	Screen Distortion
VF	Visual Flicker
VG	Visual Glitch
VR	Virtual Reality

1 Introduction

Virtual Reality (VR) technology has become a powerful tool for creating immersive educational environments. Immersion in VR is defined as a state in which individuals concentrate on the intended task and consequently ignore unrelated sensory stimuli both within and outside the virtual environment [43]. This deep engagement reduces real-world distractions and enhances learning by promoting sustained attention and improved knowledge retention [1].

Distraction in VR environments represents a significant shift in cognitive focus that disrupts user engagement and learning outcomes. In today's fast-paced digital world, where attention is increasingly fragmented and engagement levels are declining, the ability to detect and address these disruptions has become crucial, particularly in educational applications. These distractions can stem from environmental factors (e.g., background noise, visual clutter) or psychological factors (e.g., emotional stressors, shorter attention span), the latter of which are particularly challenging to detect. Such distractions can cause detrimental effects in the students' learning process [40, 12].

Despite the growing adoption of VR in educational settings, maintaining student attention and detecting distraction remains a significant challenge. While previous work has demonstrated the effectiveness of eye-tracking data in identifying external distractions, such as off-task gaze patterns in VR [41], psychological distractions — where students may appear visually engaged but are internally disengaged —remain understudied. Neurophysiological studies using EEG signals have revealed attention lapses despite outwardly focused behavior [2, 9], highlighting

the need for more sophisticated distraction detection systems.

1.1 Motivation

In regular classrooms, teachers can observe when students pay attention through eye contact, posture, and facial expressions, naturally adjusting their teaching based on these visual cues. However, in educational VR environments, these essential indicators are often hidden, and students encounter new distractions unique to virtual settings. Creating systems that detect when students are distracted in VR would be extremely valuable for helping teachers modify their teaching methods or enabling automatic adjustments to the VR environment to improve student engagement and learning outcomes.

Current VR systems have limited capability in detecting subtle shifts in attention. While some research has combined eye tracking with EEG measurements to identify distractions in controlled settings [4, 49], few models account for individual differences in how distraction manifests in brain activity. Recent reviews of brain-responsive technologies highlight this limitation [27].

Our study collects multimodal data (EEG, heart rate, eye tracking, VR movement, and HMD signals) to build machine-learning models to detect VR distractions. By considering how different students exhibit distraction in various ways, we aim to develop personalized machine learning models that adapt to each learner's unique patterns, thereby creating more effective methods for identifying attention lapses in educational virtual reality settings.

1.2 Research Questions

This work contributes to answering the following research questions:

1. How effectively can a generalized machine learning model trained on multimodal data (EEG and HMD) detect distracted students in an educational VR environment?

2. How well does the machine-learning model generalize to unseen data, such as data from different individuals or a different day?

3. Is a personalized machine-learning model better than a generalized model for detecting distracted students in a VR environment?

1.3 Contributions

This research contributes to the emerging field of Immersive Learning. This educational strategy leverages VR, AR, MR, and game-based learning to create realistic, interactive experiences that improve skill acquisition and understanding [10]. Through more accurate distraction detection, we seek to enhance the effectiveness of VR-based education and provide educators with tools to better monitor and support student engagement.

This thesis proposes a personalized, supervised machine learning framework that integrates EEG signals with head-mounted display (HMD) interaction data to detect environmental and psychological distractions in virtual reality (VR) classrooms. We collected multimodal data from 35 participants in a VR educational lecture

environment, simulating scenarios with controlled environmental distractions (e.g., visual and auditory). By training individualized long short-term memory (LSTM) models, we classify distraction states with granularity tailored to each user's neurocognitive profile. This approach builds on evidence that personalized EEG biomarkers improve engagement detection in VR [46] and aligns with calls for adaptive VR systems that respond to real-time cognitive states [27, 26].

1.4 Scope

This research aims to develop and evaluate a personalized machine-learning approach for detecting environmental and psychological distractions in virtual educational environments. Specifically, this thesis focuses on:

- Creating a multimodal dataset combining EEG signals and HMD interaction data from participants engaged in VR educational experiences
- Developing personalized LSTM models capable of detecting distractions
- Evaluating the effectiveness of personalized models compared to generalized.

1.5 Outline

The outline of the thesis is as follows:

- Chapter 2 establishes the theoretical foundation by introducing key concepts and terminology related to distractions while reviewing the existing literature in this field.

- Chapter 3 details our methodological approach, including data collection and processing procedures, training and testing protocols, and provides information about the datasets utilized in this research.
- Chapter 4 presents our experimental framework, comprehensive results, and analysis. This chapter includes the evaluation metrics employed and demonstrates the performance outcomes of our approach.
- Chapter 5 concludes the thesis by summarizing our contributions, acknowledging limitations, and proposing directions for future research.
- Chapter 6 includes appendices with supplementary materials on window length parameters, comparison graphs for all research questions, and participant-specific analysis.

2 Related Work

This chapter reviews previous research on distraction detection in immersive educational VR environments. The following sections summarize key contributions in this field.

Behavioral Indicators of Engagement and Distraction

Reduced head movement has been shown to strongly correlate with higher self-reported immersion and engagement, as participants who exhibited less head movement rated the VR content as more engaging [30]. Similarly, eye gaze patterns have been associated with the state of immersion [8], and visual cues in VR teaching environments enable educators to identify distracted students [7] more efficiently.

Investigations of gaze transition patterns suggest that changes in fixation duration and transitions can serve as reliable indicators of user engagement. These measures could enhance models for detecting and mitigating attentional lapses in immersive VR settings [36]. Complementary research has demonstrated that pupil dilation can predict cognitive failure; by integrating real-time eye-tracking data, our model aims to detect cognitive overload and anticipate lapses in attention [37].

Immersion and Presence in VR

A robust sense of immersion, presence, agency, usability, and overall user experience appears to mitigate distractions within VR settings [5]. The concept of immersion is multifaceted, as it can be understood both as a psychological state and as a technological property of VR systems that facilitates engaging experiences.

Immersion often involves a shift in the attentional state, wherein cognitive processes become so absorbed by the VR experience that the user becomes less aware of the physical environment [1].

While immersive learning environments can significantly enhance engagement, evidence suggests poorly structured strategies may induce cognitive overload or distractions. Well-organized content and coherent teaching strategies are essential to minimize these issues and help maintain focus [6].

A strong sense of presence in VR, measured through subjective and objective metrics, is linked to improved task performance and fewer distractions, suggesting that real-time monitoring of presence could be valuable in distraction detection systems [50]. Auditory presence studies using loudspeaker setups have found correlations between subjective questionnaire ratings of presence and objective EEG measures, further affirming the utility of EEG in measuring immersive presence or distractions [51].

EEG and Physiological Monitoring

EEG has emerged as a powerful method for monitoring dynamic brain states. Studies indicate that higher engagement is accompanied by increased neural synchrony, attention, and emotional involvement [11]. Additionally, meta-analyses have confirmed that immersive VR can effectively reduce pediatric pain and anxiety through distraction, with success measurable via physiological biomarkers (EEG, heart rate) and behavioral markers (eye-tracking). These findings support the development of adaptive VR content that responds in real-time to individual stress

responses and interaction patterns [14]. EEG-based distraction detection, which relies on identifying brainwave patterns that signal cognitive overload or attention shifts, remains highly accurate but is typically considered intrusive and primarily used in research settings [21]. Research also shows that specific brain responses to visual stimuli occur predictably in terms of timing and location [29] and that immersive environments can significantly modulate cognitive focus, an essential factor when assessing attentional shifts in VR-based studies [35].

Psychological distractions correlate with reduced cognitive performance and retention [53]. For instance, EEG studies show that they disrupt frontal theta synchronization and parietal alpha desynchronization, biomarkers of working memory and visuospatial processing [53, 25]. Research into EEG signal variations during problem-solving tasks reveals that different levels of mental effort correspond with specific changes in EEG frequency bands, offering further insight into cognitive load management in VR [53].

Intentional Distraction Studies

The intentional integration of distractions in VR environments provides a controlled method for studying their effects on learning, attention, and engagement. Researchers can assess how learners respond to distractions by simulating real-world interruptions, such as phone calls, door knocks, or background noise, and develop strategies to enhance cognitive resilience. This approach is particularly promising in training scenarios where managing distractions is critical, such as in pilot training, emergency response, or high-stress decision-making [47]. Research

also indicates attentional shifts occur when users divert their focus from VR to external disturbances [38, 15, 44].

Furthermore, effective distraction detection must consider both external interruptions and internal cognitive factors [45]. Interestingly, some studies have leveraged well-timed distractions as a positive intervention to reduce cybersickness by diverting attention from discomfort, thereby underscoring that controlled distractions can be beneficial under certain conditions [48].

Cross-Subject Validation and Personalization

Recent work on cross-subject EEG validation and multimodal feature fusion has improved emotion recognition accuracy by enhancing model generalizability despite ongoing challenges with inter-subject variability. While current intra-subject EEG-based emotion recognition methods achieve high accuracy (96%) after extensive training (over 30 minutes per user), cross-subject approaches typically reach only about 58% accuracy, highlighting a critical bottleneck for practical applications [42, 19, 13, 52, 24, 22, 32, 34, 20, 33, 31].

Summary

We discuss some of the recent approaches relevant to our work. Literature reveals significant research into the detection of distractions within VR educational environments. While prior studies employ single-modal approaches, head movement [30], eye gaze patterns [8], and EEG measurements [11, 21], some researchers have begun exploring multimodal methods. Our work advances these multimodal

approaches by integrating a comprehensive array of physiological and behavioral signals, including EEG, heart rate, eye tracking, VR movement, and head-mounted display data, to develop more robust machine learning models. These models recognize distractions while effectively accounting for individual differences among learners. We introduce a personalized approach by expanding upon cross-subject validation studies [42, 19, 13, 52, 24, 22, 32, 34, 20, 33, 31].

3 Experimental Setup and Methods

This chapter describes our methodology, including the educational VR environment and our user study experiment, which was conducted to collect data for the machine learning model. It describes the design and implementation of our physics-based immersive learning environment, which incorporates controlled audio and visual distractions to simulate real-world interruptions. The chapter outlines our multimodal data collection approach, capturing EEG signals, eye-tracking metrics, VR movement data, and heart rate measurements from participants as they engage with educational content in VR. It also explains our data preprocessing methods, synchronization techniques, and the development of LSTM-based models for detecting distraction using three distinct strategies: generalized, participant-independent, and personalized approaches.

3.1 Participants

A total of 35 participants (22 male, 13 female) from the university population participated in this study, with an average age of 25.69 years (SD = 6.6). All subjects had normal or corrected-to-normal vision. Subjects who opted out of EEG data collection or had poor-quality EEG recordings were excluded, resulting in a final dataset of 27 subjects. Of these, 5 volunteers agreed to participate in a 2nd data collection session, conducted 2–3 days after their initial session, specifically for training and testing the personalized model. The study was approved by the IRB of the University of Louisiana at Lafayette and proceeded only after subjects completed vulnerability screening questionnaires and provided signed informed consent.

3.2 Experiment

A VR-based educational environment was designed to teach projectile motion through an immersive physics lecture. The scene featured a lecturer explaining key concepts while narrating a live demonstration. During the demonstration, a cannon was fired to visually illustrate the effects of gravity on a projectile's trajectory. After the lecture, participants engaged in an interactive activity, applying their learning by controlling the cannon themselves in a simulated target practice, reinforcing their understanding through hands-on experience.

3.3 Experiment Materials

During the experiment, participants were exposed to various distraction stimuli, categorized as audio and visual distractions. Simultaneously, data were collected from EEG, HMD, and HR devices. The following section provides a detailed description of the distraction interventions and the recordings of these data.

3.3.1 Distraction Interventions

Nine distinct distraction types were introduced, designed to simulate real-world interruptions and VR-specific anomalies. These controlled distractions were introduced using a Latin square design, ensuring that randomized sequences were assigned to each participant. Table 1 provides an overview of these distraction types. Figure 1 shows the visual distractions: Visual Glitch, Screen Distortion, Content Anomaly, and Visual Flicker.

Audio Distractions	
External Knock	A simulated knocking sound designed to mimic an unexpected interruption.
Phone Alert	A ringtone or notification sound simulating a mobile phone disturbance.
Ambient Noise	Background sounds, such as chatter or construction noise, make it difficult to focus.
Audio Dropout	Simulated intermittent loss of sound, replicating poor connectivity issues.
Echo Effect	A reverberation effect applied to speech, making it harder to understand.
Visual Distractions	
Visual Glitches (VG)	Distorted visuals resembling VR display errors, such as pixelation or artifacts.
Screen Distortion (SD)	Sudden shifts in object shapes, colors, or screen movement, creating visual confusion.
Content Anomaly (CA)	Appearance of the out-of-context element, such as the mushroom of a Mario game in our VR lecture.
Visual Flicker (VF)	Rapidly appearing and disappearing objects causing distractions in the scene.

Table 1. Categories of Distractions in the VR Environment

3.3.2 Devices and Equipment

The experimental setup utilized multiple devices to capture neural, physiological, and behavioral data. EEG signals were recorded using a 16-channel EEG cap from OpenBCI, with all electrodes connected to the Cyton board. The board transmitted data via Bluetooth to a PC, where signals were logged using the OpenBCI GUI at a sampling rate of 125 Hz.

Participants wore an HTC Vive Eye Pro HMD for VR interactions, recording eye-tracking data at a rate of 90 Hz. Polar Verity Sense for heart rate (HR) at 4 Hz.

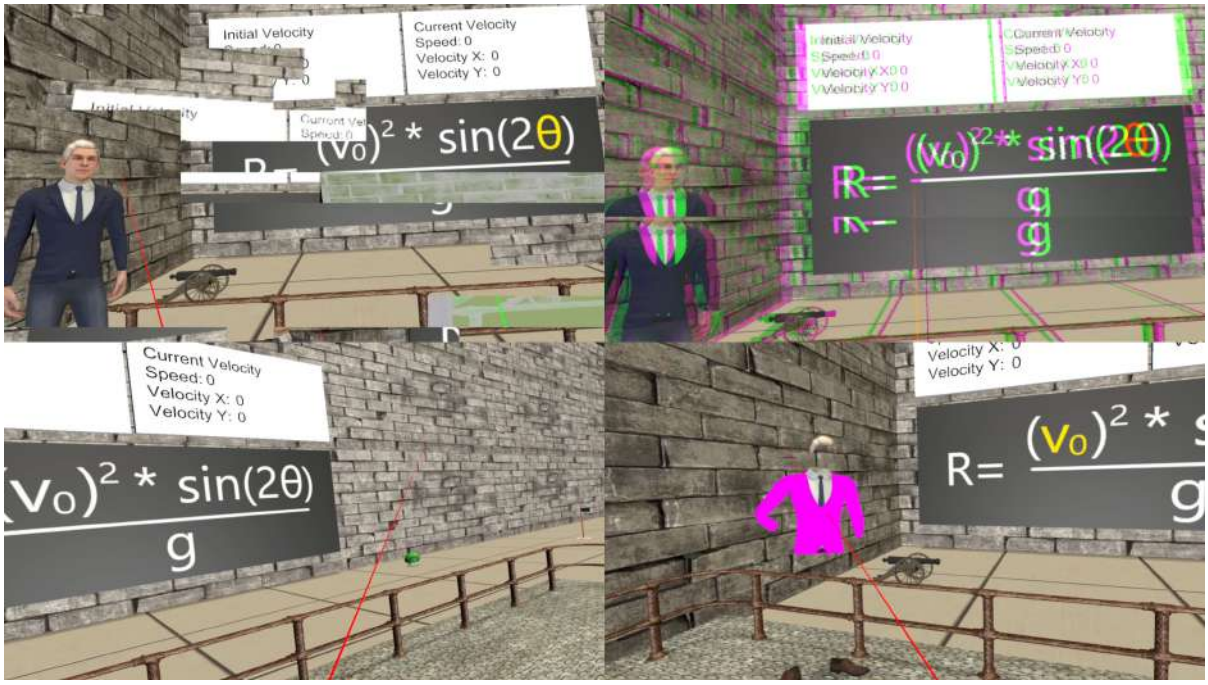


Figure 1. Visual Distractions (from left to right): Visual Glitch, Screen Distortion, Content Anomaly, and Visual Flicker.

VR game data were recorded at 90 Hz to track real-time user interactions.

Model training and data processing were conducted on a high-performance desktop system featuring an Intel Core i9-11900F processor, Windows 10 Pro, an NVIDIA GeForce RTX 4080 GPU, and 64 GB of RAM. Data analysis and machine learning models were implemented using Python 3.9.14 and relevant libraries, including PyTorch, NumPy, pandas, and scikit-learn.

3.4 Experiment Protocol and Setup

Each experimental session lasted approximately 40 minutes, including the consent process, device setup, and calibration. Participants engaged in a VR lecture on projectile motion, which was delivered through a combination of audio narration, slides, animations, and a teacher avatar. The lecture was divided into ten sections,

with nine sections incorporating distractions while one section remained distraction-free as a control. Following the lecture, participants applied their learning by interacting with a virtual cannon, aiming, and firing projectiles in the VR environment.

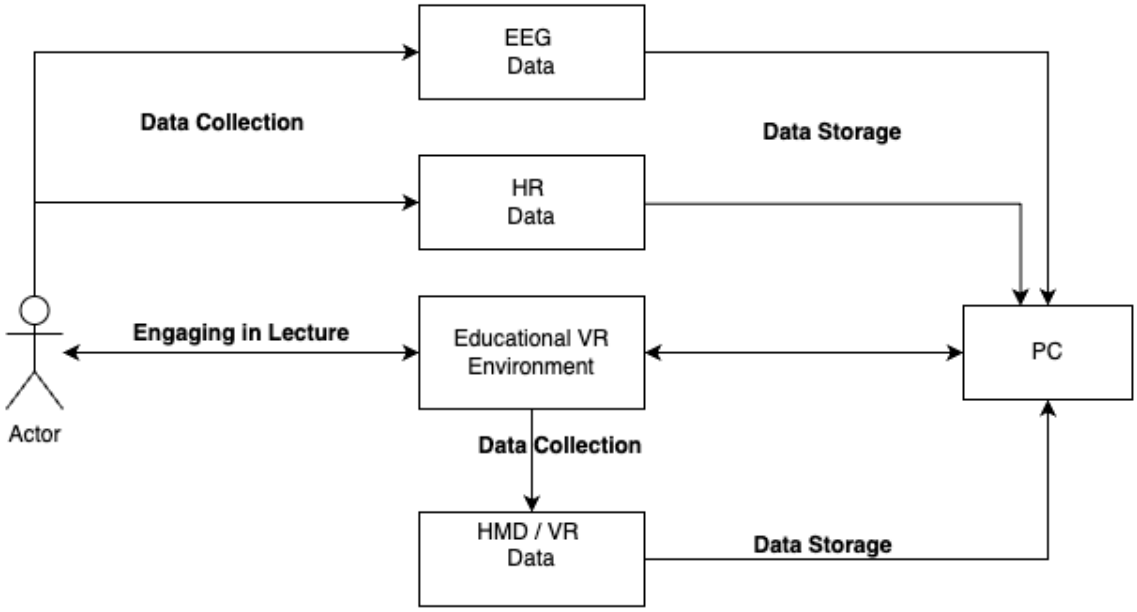


Figure 2. Experiment Procedure

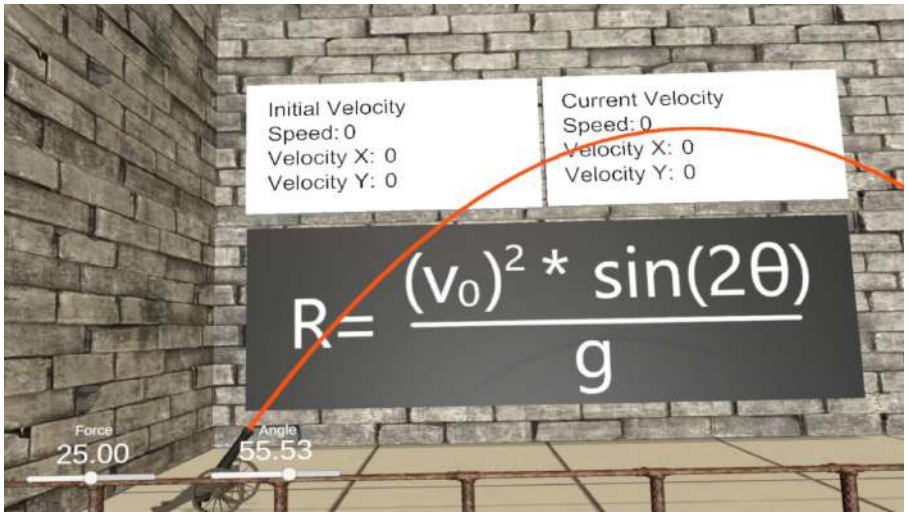


Figure 3. Educational Environment

(a) Participant **(b)** Participant

Figure 4. Experimental Setup

3.5 EEG Data

EEG data were collected using an OpenBCI all-in-one electrode cap with 16 channels. The signals were recorded at a sampling frequency of 120 Hz. The electrode placements followed a standard montage and included FP1, FP2, F3, Fz, F4, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, O1, and O2. The EEG data were stored in a time series.

3.6 HMD Data

Eye-tracking and VR movement data were captured using the HTC Vive Eye Pro headset. The eye-tracking data included left and right pupil diameter (in mm), eye openness for both eyes, gaze direction, gaze origin, game gaze direction, and pupil position. The VR movement data collected from the HMD encompassed the headset forward vector, representing its orientation in 3D space, and the headset velocity,

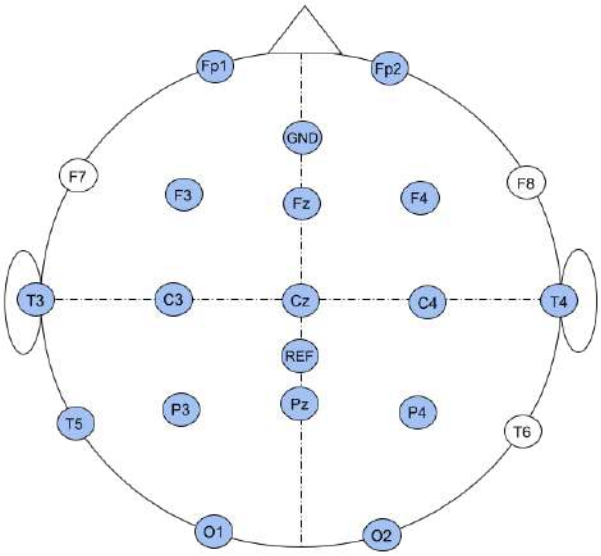


Figure 5. EEG electrode placements in standard 10-20 montage

which tracks its movement dynamics. Additionally, the focus point and focus normal were recorded to determine the exact surface and angle of the user's gaze. The right controller's position, rotation, velocity, acceleration, angular velocity, and angular acceleration were also captured to track hand movements and interactions in the VR environment. These datasets were logged in sync. Details of these data categories are provided in the Tables 2 and 3

3.7 Synchronization

Since EEG and VR data were recorded at different sampling rates (125 Hz for EEG and 90 Hz for VR), a custom synchronization process was implemented to ensure accurate alignment. Forward fill interpolation was used to align the most recent EEG value with each Unity timestep, and timestamps were synchronized with the VR data. Both data streams were then merged into a single, time-synchronized CSV file per participant, ensuring that all recorded distraction events were accurately

Heart Rate and Eye Tracking	
HeartRate	Heart rate measurement in BPM
LeftEyeOpenness, RightEyeOpenness	Eye openness level for each eye
CombinedEyeOpenness	Average openness of both eyes
LeftPupilDiameter, RightPupilDiameter	Pupil diameter in mm
Combined Gaze Convergence Distance	Convergence distance between both eyes in mm
Focus and Object Interaction	
FocusDistance	Distance to the object currently in focus
FocusColliderObject	The VR object being observed
FocusRigidbody	Physical properties of the observed object
FocusTransform	Transform properties of the focused object
Gaze Direction and Origin	
LeftGazeDirection(X, Y, Z), RightGazeDirection(X, Y, Z)	Normalized direction vectors of each eye's gaze
CombinedGazeDirection (X, Y, Z)	Combined gaze direction of both eyes
LeftGazeOriginMM(X, Y, Z), RightGazeOriginMM(X, Y, Z)	3D coordinates of gaze origin in mm
CombinedGazeOriginMM (X, Y, Z)	Combined gaze origin from both eyes
Pupil Position	
LeftPupilPosition(X, Y), RightPupilPosition(X, Y)	Screen-space coordinates of each pupil
CombinedPupilPosition (X, Y)	Average pupil position of both eyes

Table 2. Heart Rate, Eye Tracking, Focus, and Gaze Data in VR Experiment

aligned across modalities. This synchronization enabled precise analysis and effective model training.

VR Environment Gaze and Movement	
GameGazeDirection (X, Y, Z)	Gaze direction relative to the game environment
HeadsetForwardVector (X, Y, Z)	Orientation vector of the VR headset
HeadsetVelocity (X, Y, Z)	Speed and movement of the headset
Focus Point and Surface Normal	
FocusPoint (X, Y, Z)	3D coordinates of the user's focus point
FocusNormal (X, Y, Z)	Normal vector of the surface the user is looking at
Controller (Right Hand) Position and Rotation	
RightPosition (X, Y, Z)	3D position of the right-hand controller
RightRotation (X, Y, Z, W)	Quaternion representation of controller rotation
RightVelocity (X, Y, Z)	Velocity of the right controller
RightAcceleration (X, Y, Z)	Acceleration of the right controller
RightAngularVelocity (X, Y, Z)	Angular velocity of the right controller
RightAngularAcceleration (X, Y, Z)	Angular acceleration of the right controller

Table 3. VR Environment, Focus Point, and Controller Data in VR Experiment

3.8 Dataset and Labelling

To establish labeled training data for the machine learning model, participants were asked three self-report questions at predefined intervals in the lecture:

1. Presence of Distraction: "Since the last pause, have you experienced any distractions, whether from the lecture, the environment, or your own thoughts (e.g., mind wandering)?" (Bool)

2. Intensity of Distraction: "If yes, on a scale from 1 to 7, how would you rate your level of distraction, where 1 is 'barely distracted' and 7 is 'highly distracted'?" (Integer)

3. Nature of Distraction: "Please specify the nature of your distraction: Was it related to the lecture, the environment, your own thoughts, or something else?" (String)

Responses were recorded and used to label EEG and VR data, marking timestamps of reported distractions. The baseline session was labeled as an undistracted baseline with no distractions.

3.9 Data Preprocessing and Preparation

The preprocessing step was crucial in preparing EEG and VR data for machine learning-based distraction classification. It began by iterating through multiple CSV files containing recorded participant data and systematically applying a series of transformations to clean, filter, and structure the dataset. The 2D and 3D, 4D array fields that had coordinates values were transformed in their respective coordinate fields, for example, LeftGazeDirection was transformed to LeftGazeDirection_X and LeftGazeDirection_Y, RightGazeOriginMM was transformed to RightGazeOriginMM_X, RightGazeOriginMM_Y, RightGazeOriginMM_Z, RightRotation was transformed to RightRotation_X, RightRotation_Y, RightRotation_Z, RightRotation_W.

Columns distraction level, interval, sample index, FocusColliderObject, FocusTransform, FocusRigidbody, CombinedEyeOpenness, CombinedGazeConvergenceDistanceMM, CombinedPupilPosition,

RightAcceleration, RightAngularAcceleration, and HeartRate, were removed. Some of these columns consisted entirely of zeros or constant values throughout the experiments. At the same time, text-based fields such as FocusRigidbody and FocusColliderObject were discarded due to their incompatibility with numerical analysis.

The sliding window approach was applied to generate fixed-length sequences, accommodating the temporal nature of EEG and VR data and ensuring that temporal dependencies are preserved. The sequences were created with a 50% overlap, allowing the model to learn from overlapping contextual information and improving classification accuracy. Each generated sequence was labeled based on the distraction state at the last time step, maintaining chronological integrity in the dataset. Additionally, sequence-level distraction counts are updated per participant, allowing for the analysis of how distractions are contained over time.

Finally, the processed sequences and corresponding labels were downstreamed to model training by doing an 80-20 split on the sequenced data. By structuring the data efficiently and ensuring alignment between EEG and VR signals, the preprocessing pipeline enabled robust training of deep learning models, using LSTM, for distraction classification in immersive VR environments

3.10 Model Architecture

Long Short-Term Memory (LSTM) networks, a specialized variant of Recurrent Neural Networks (RNNs), are designed to mitigate the vanishing gradient problem inherent in traditional RNNs. LSTMs effectively capture long-term dependencies in

sequential data, including time series, text, speech, and EEG signals [18, 16] Through a gated mechanism that regulates the flow of information. LSTM models have demonstrated high classification accuracy in recognizing emotional states Their ability to retain both distant and recent events makes them particularly effective for emotion prediction, as past EEG activations significantly influence the target predictions, providing valuable insights into underlying emotional states [3]

The LSTM architecture employed in this study consisted of 256 or 512 nodes per layer, with four stacked LSTM layers. A single Dense layer with one node was included, followed by a sigmoid activation function. Dropout rates of 20%, 30%, or 50% were applied between the LSTM layers to prevent overfitting. [23] demonstrated that the Adam optimizer is a more suitable optimization method for complex neural networks.

The model used a comprehensive training strategy, incorporating carefully selected initialization methods, loss functions, and optimization techniques to enhance stability and performance.

Xavier uniform initialization is applied to the LSTM's input-to-hidden weights for weight initialization. In contrast, orthogonal initialization is used for hidden-to-hidden weights, with bias terms set to zero. This initialization scheme facilitates stable gradient flow throughout the network [17] Empirical studies have demonstrated that orthogonal initialization significantly improves training stability in LSTMs [28]

The training process utilizes Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss), incorporating class weighting to address potential class imbalances in the dataset. Class weights are computed using a balanced strategy,

with specific adjustments for the positive class and a fallback to 1.0 in single-class scenarios.

For optimization, the AdamW optimizer is employed with configurable learning rates (0.001 or 0.0001) and a weight decay of $1e^{-5}$ for L2 regularization. To further enhance training stability, gradient clipping is applied with a maximum norm of 1.0 [28, 39]

Training is conducted with a batch size of 128 for up to 100 epochs, incorporating an early stopping mechanism with a patience of 10 epochs. Additionally, a ReduceLROnPlateau scheduler monitors validation loss and reduces the learning rate by a factor of 0.1 when training stagnates, using the same patience period as early stopping. This integrated approach ensures robust training and effective model convergence, mitigating challenges such as class imbalance and learning rate fluctuations.

3.11 Training / Testing Strategies

An LSTM model was created, and k-stratified validation was performed over 10 folds. Early stopping was implemented to prevent overfitting, and hyperparameters were systematically tuned to achieve optimal performance. The sequence lengths and hyperparameter values were explored, including dropout rate, hidden layer dimensions, number of layers, and learning rate.

The function conducted hyperparameter tuning for each sequence length by generating all possible combinations of the specified parameters and iterating through them. The best-performing hyperparameters were tracked based on

validation loss and stored for reference during the training process. Once the optimal hyperparameters for a given sequence length were identified, the model was fine-tuned by retraining it using these settings.

The dataset was split into an 80%-20% training-validation set and a 20% test set, ensuring that class distribution was preserved through stratification. A 10-fold stratified cross-validation (StratifiedKFold) was applied to the training-validation split to systematically assess model performance across different data subsets. Training and validation data were separated for each fold, and input features were standardized using StandardScaler. The LSTM model was then instantiated with the specified hyperparameters, transferred to the appropriate device (CPU/GPU), and initialized with optimized weight initialization strategies to enhance convergence.

Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) was used as the loss function, with computed class weights to effectively handle class imbalances. If only one class was present in the training data, a default weight was assigned to prevent errors. The AdamW optimizer and a learning rate scheduler (ReduceLROnPlateau) were configured to dynamically adjust the learning rate based on the validation loss.

The training was performed in batches, initially with 100 epochs chosen for efficient processing. Each epoch involved forward passes through the LSTM model, loss computation, and backpropagation with gradient clipping to prevent the gradients from exploding. Validation was conducted at the end of each epoch, and the learning rate was adjusted accordingly. An early stopping mechanism with a patience of 10 epochs was implemented to halt training if no improvement in validation loss was

observed. After training, the model was evaluated on the validation set, with key performance metrics including accuracy, precision, recall, F1 score, and AUC calculated. The best model state (weights) was saved if it achieved the highest validation accuracy across folds. The validation confusion matrix was also accumulated across all folds to provide a comprehensive view of classification performance.

Once cross-validation was complete, the model was evaluated on the held-out test set. The test data was standardized using the same scaler as the training data, and the best-performing model was loaded for inference. Predictions were generated, and performance metrics were computed, including ROC-AUC, PR-AUC, and F1 score at the optimal threshold. The final test confusion matrix was also plotted to visualize classification performance.

At the end of the process, the mean cross-validation scores, test performance metrics, and best hyperparameter configuration were saved. The trained model's state dictionary was also stored to ensure reproducibility and facilitate further usage. The process systematically explored different sequence lengths and hyperparameter configurations, selecting the best combination based on validation loss. The final model was created using the most effective parameters, ensuring optimal performance.

3.11.1 Generalized Model

A generalized model was developed using the complete dataset, using an 80%-20% split between training and testing sets. All participant data were pooled

together. This approach aimed to develop a universal distraction detection model that can generalize across different individuals. Specifically, the training portion (80% of the complete dataset) was used for model development and hyperparameter tuning, while the testing portion (remaining 20%) was reserved exclusively for final performance evaluation of model generalizability.

3.11.2 Participant-Independent Models

For the participant-independent approach, a leave-one-out cross-validation strategy was implemented. For each model, data from $n-1$ participants constituted the training set, where n represents the total number of participants, while data from the excluded participant formed the test set. Within the training set containing $n-1$ participants, an 80%-20% split was further applied to create training and validation subsets. This procedure was systematically repeated n times, generating a suite of n distinct models, with each participant serving as the exclusive test case exactly once. This approach assessed the model's ability to generalize to entirely unseen individuals, providing crucial insights into its potential real-world applicability across diverse users without requiring person-specific calibration.

3.11.3 Participant-Specific/Personalized Models

The participant-specific models were constructed on a per-individual basis, leveraging the temporal separation between data collection sessions. For each participant, two complementary models were developed: (1) a model trained on the first session data (with an internal 80%-20% training-validation split) and tested on

the complete second session data, and (2) a model trained on the second session data (with an internal 80%-20% training-validation split) and tested on the complete first session data. This cross-session validation strategy evaluated the temporal stability of distraction patterns within individuals and assessed whether personalized models could effectively capture user-specific characteristics of distraction that persist across different time periods.

4 Results and Discussion

4.1 RQ1: Generalized Model

The model underwent rigorous training and testing across various sequence lengths/window lengths (5, 10, 15, 20, 25, 30, 35, 40, 50, 75) and multiple parameter configurations, including dropout rates (0.2, 0.3, 0.5), hidden dimension sizes (256, 512), number of layers (4), and learning rates (0.001, 0.0001). Hyperparameter tuning was performed to identify the optimal configuration and determine the best-performing model.

The best configurations and hyperparameters for each sequence are shown in Table 4.

ID	Seq Length	Dropout	Hidden Size	Layers	Learning Rate
C1	5	0.5	256	4	0.0001
C2	10	0.5	256	4	0.0001
C3	15	0.5	512	4	0.0001
C4	20	0.5	512	4	0.0001
C5	25	0.3	256	4	0.0001
C6	30	0.5	256	4	0.0001
C7	35	0.5	256	4	0.0001
C8	40	0.5	256	4	0.0001
C9	50	0.2	256	4	0.0001
C10	75	0.5	256	4	0.0001

Table 4. Model Configuration Mapping

ID	Cross Validation Accuracy		Cross Validation Loss			Test Performance					
	Train	Validation	Train	Val	Test	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC
C1	0.844	0.841	0.435	0.436	0.402	0.930	0.951	0.929	0.974	0.900	0.960
C2	0.829	0.827	0.440	0.441	0.424	0.864	0.902	0.906	0.899	0.840	0.938
C3	0.823	0.818	0.442	0.444	0.433	0.854	0.897	0.885	0.910	0.817	0.929
C4	0.802	0.792	0.448	0.452	0.446	0.815	0.867	0.872	0.863	0.783	0.915
C5	0.816	0.810	0.443	0.446	0.433	0.848	0.892	0.889	0.894	0.816	0.929
C6	0.804	0.793	0.447	0.451	0.459	0.783	0.843	0.852	0.834	0.748	0.901
C7	0.798	0.793	0.449	0.451	0.436	0.838	0.885	0.884	0.886	0.806	0.925
C8	0.799	0.790	0.448	0.453	0.443	0.825	0.876	0.872	0.879	0.788	0.918
C9	0.787	0.771	0.454	0.461	0.466	0.777	0.843	0.831	0.855	0.724	0.894
C10	0.776	0.758	0.458	0.467	0.465	0.775	0.840	0.837	0.844	0.728	0.895

Table 5. Performance Metrics for General Models

Optimal performance was achieved using a sequence length/window length of

5, with the following configuration: dropout (0.5), hidden size (256), number of layers (4), and learning rate (0.0001), as determined by evaluating the test metrics as shown in Table 5. The model attained an accuracy of 0.930, precision of 0.929, recall of 0.974, F1 score of 0.951, ROC AUC of 0.900, and PR AUC of 0.960. Additionally, the balance score was 0.9163, with a mean score of 0.9406 and a standard deviation of 0.0243, highlighting its strong and consistent performance across evaluations.

4.2 RQ2: Participant Independent Models

In this approach, the leave-one-out strategy was used. The model was trained on data from 26 participants and tested on the 27th. During training, the model was unable to access the participants' data being tested. This process ensures the evaluation of generalization across subjects. For instance, in the first iteration, the model is trained on data from 26 participants, excluding subject 1, and is then tested solely on subject 1's data. In the second iteration, a new model is trained on all data except that of subject 2 and then tested exclusively on subject 2. This process is repeated for all 27 participants, resulting in 27 models, each of which is tested for its ability to generalize across different individuals.

Since optimal parameters were identified in RQ1, they were used for training. The models underwent rigorous training and testing across the following parameters: sequence length (5), dropout rate (0.5), hidden dimension sizes (256, 512), number of layers (4), and learning rate (0.0001).

The training and validation results are shown in Table 6, and performance on unseen data is shown in Table 7.

	Accuracy		Loss			Test Performance					
PID	Train	Val	Train	Val	Test	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC
5	0.905	0.904	0.402	0.403	0.391	0.933	0.954	0.937	0.971	0.905	0.964
6	0.906	0.904	0.412	0.413	0.404	0.928	0.950	0.924	0.977	0.895	0.959
7	0.922	0.920	0.409	0.410	0.403	0.932	0.952	0.934	0.971	0.906	0.962
8	0.924	0.921	0.405	0.406	0.395	0.940	0.958	0.943	0.973	0.918	0.967
9	0.905	0.903	0.411	0.412	0.404	0.924	0.947	0.926	0.968	0.894	0.958
10	0.897	0.896	0.419	0.419	0.412	0.916	0.941	0.917	0.967	0.883	0.953
11	0.910	0.908	0.413	0.414	0.402	0.933	0.953	0.935	0.972	0.909	0.963
12	0.918	0.916	0.411	0.412	0.404	0.930	0.951	0.932	0.971	0.904	0.961
13	0.922	0.919	0.411	0.413	0.406	0.932	0.952	0.932	0.973	0.907	0.962
14	0.908	0.907	0.406	0.407	0.396	0.925	0.948	0.937	0.959	0.901	0.962
17	0.908	0.906	0.413	0.414	0.404	0.931	0.951	0.928	0.976	0.902	0.960
18	0.906	0.905	0.415	0.416	0.406	0.934	0.953	0.929	0.979	0.905	0.961
19	0.911	0.908	0.413	0.414	0.400	0.940	0.958	0.936	0.980	0.914	0.965
20	0.902	0.900	0.416	0.417	0.405	0.930	0.951	0.926	0.978	0.899	0.960
21	0.917	0.915	0.398	0.400	0.390	0.931	0.953	0.934	0.973	0.900	0.963
22	0.899	0.896	0.419	0.421	0.403	0.938	0.956	0.935	0.978	0.913	0.964
24	0.906	0.904	0.407	0.408	0.400	0.925	0.948	0.923	0.975	0.889	0.958
25	0.890	0.889	0.416	0.416	0.399	0.928	0.950	0.931	0.969	0.899	0.961
26	0.906	0.905	0.416	0.416	0.402	0.939	0.957	0.937	0.979	0.915	0.965
27	0.891	0.890	0.412	0.412	0.398	0.929	0.951	0.926	0.977	0.894	0.960
28	0.905	0.904	0.413	0.414	0.398	0.938	0.956	0.944	0.969	0.918	0.967
29	0.915	0.913	0.385	0.386	0.379	0.933	0.955	0.941	0.969	0.903	0.966
30	0.897	0.894	0.420	0.422	0.406	0.928	0.949	0.926	0.974	0.898	0.959
31	0.918	0.916	0.406	0.407	0.392	0.946	0.962	0.948	0.977	0.926	0.970
39	0.916	0.912	0.411	0.413	0.402	0.933	0.953	0.936	0.971	0.909	0.964
40	0.900	0.898	0.420	0.421	0.404	0.934	0.953	0.934	0.973	0.910	0.963
41	0.901	0.900	0.416	0.417	0.411	0.919	0.943	0.915	0.974	0.883	0.953

Table 6. Cross-Validation Performance Metrics

PID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC	Loss
5	0.407	0.483	0.379	0.667	0.544	0.604	0.811
6	0.700	0.824	0.700	1.000	0.500	0.850	0.550
7	0.803	0.868	0.851	0.885	0.677	0.914	0.420
8	0.747	0.661	0.796	0.566	0.649	0.820	0.509
9	0.707	0.131	0.255	0.088	0.129	0.533	0.649
10	0.800	0.434	0.596	0.341	0.191	0.726	0.577
11	0.892	0.936	0.880	0.998	0.738	0.942	0.386
12	0.823	0.888	0.834	0.950	0.706	0.928	0.423
13	0.900	0.601	0.998	0.430	0.701	0.960	0.415
14	0.500	0.659	0.496	0.983	0.492	0.743	0.770
17	0.808	0.891	0.804	1.000	0.529	0.904	0.440
18	0.900	0.904	0.892	0.916	0.451	0.940	0.356
19	0.800	0.889	0.800	1.000	0.500	0.900	0.443
20	0.801	0.887	0.803	0.990	0.529	0.901	0.441
21	0.545	0.117	0.662	0.064	0.594	0.398	0.632
22	0.900	0.042	0.956	0.021	0.696	0.956	0.515
24	0.512	0.612	0.485	0.829	0.505	0.699	0.752
25	0.642	0.755	0.606	1.000	0.556	0.808	0.653
26	0.900	0.064	0.716	0.034	0.243	0.841	0.519
27	0.500	0.042	0.403	0.022	0.345	0.413	0.599
28	0.852	0.641	1.000	0.471	0.885	0.974	0.434
29	0.000	0.000	0.000	0.000	0.000	0.500	1.297
30	0.801	0.795	0.786	0.805	0.597	0.892	0.469
31	0.706	0.823	0.700	0.999	0.521	0.852	0.551
39	0.800	0.452	0.682	0.338	0.329	0.766	0.541
40	0.900	0.937	0.900	0.976	0.528	0.951	0.339
41	0.800	0.217	0.470	0.142	0.123	0.669	0.585

Table 7. Performance Metrics on Unseen Data

The F1-score is especially critical for distraction detection, as it balances precision and recall with the system's ability to avoid false alarms while minimizing the risk of missing true distraction events. Both missing true distractions (low recall) and falsely detecting non-distractions (low precision) in VR worlds can be costly, degrading the user experience. So, F1-score is the most appropriate metric for the overall system.

The F1-score analysis reveals substantial variability in cross-session generalization. These distributions in Table 8 and Figure 6 highlight a critical challenge.

Metric	$\geq$ **0.6**	$\geq$ **0.7**	$\geq$ **0.8**	$\geq$ **0.9**
F1-Score	17 (63.0%)	12 (44.4%)	10 (37.0%)	3 (11.1%)
Precision	19 (70.4%)	15 (55.6%)	10 (37.0%)	4 (14.8%)
Recall	15 (55.6%)	14 (51.9%)	14 (51.9%)	11 (40.7%)

Table 8. Distribution of Performance Metrics on Unseen Data

While 63% of the models have relatively good performance (F1 $\geq$ 0.6), fewer than half maintain good detection capability (F1 $\geq$ 0.7) on unseen data, and only a small percentage (11.1%) have excellent generalization (F1 $\geq$ 0.9). Such a high drop-off indicates the difficulty in maintaining consistent pattern recognition (EEG patterns, Eye gaze patterns, Movement patterns) across different recording sessions.

Examining individual cases provides further insights. Participant 11 demonstrates exceptional generalization, achieving an F1-score of 0.936 on unseen data, which approaches its cross-validation performance. Similarly, Participants 18 and 40 maintain F1-scores above 0.9, suggesting their signatures of distraction

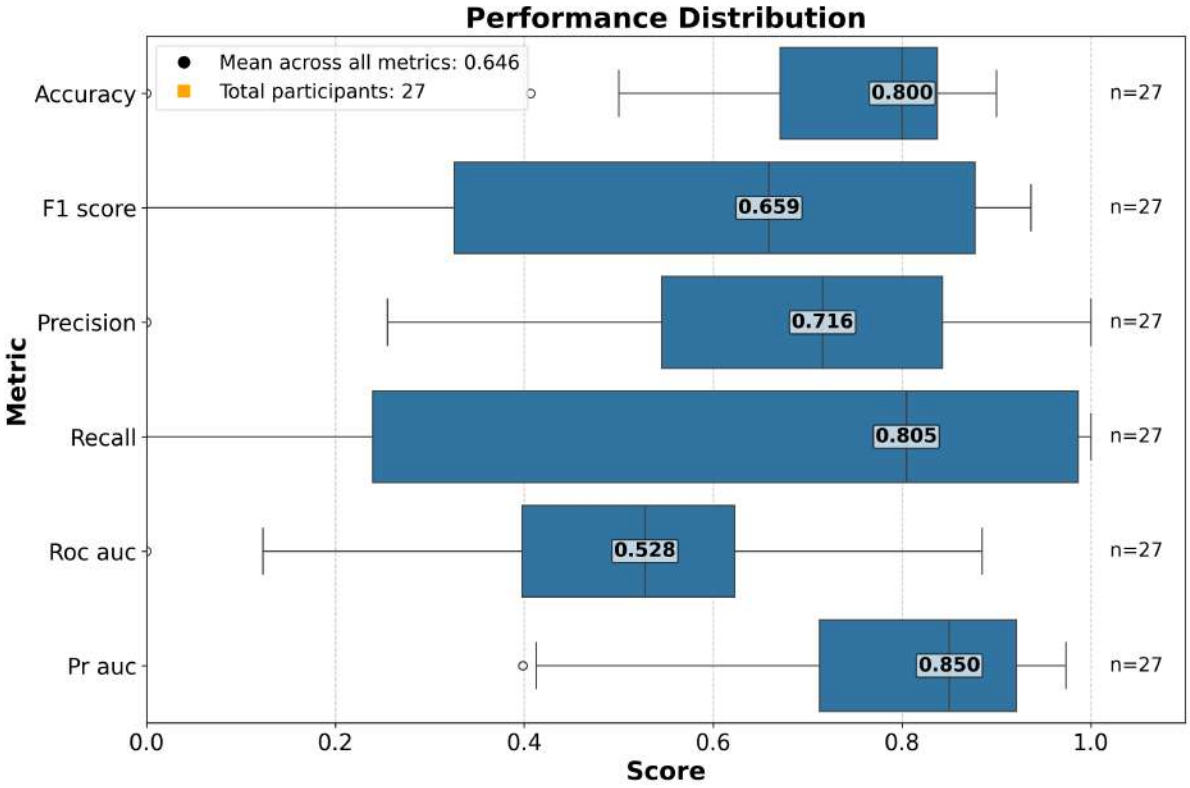


Figure 6. Distribution of Performance Metrics Across All Models

remain remarkably stable across sessions. These high-performing outliers indicate that robust cross-session distraction detection is achievable, albeit challenging.

However, despite impressive cross-validation scores, a few people (22, 26, 27, 29) exhibit degradation in F1-scores below 0.1 on unseen data. These models not only make more mistakes on fresh sessions but also make mistakes that seriously compromise the balance between precision and recall, as evidenced by the F1-score degradation, which is more severe than the losses in raw accuracy. This implies that the statistical characteristics of the distraction signatures change entirely from one session to the next.

The gap between cross-validation F1-scores (consistently high, ranging from 0.94 to 0.96) and unseen data F1-scores (highly variable, ranging from 0.042 to

0.937) represents a challenge for distraction detection systems. This disparity in F1-score performance is problematic because it indicates that standard validation procedures may give misleadingly optimistic estimates of real-world system reliability.

These results show that while models effectively detect distraction events (with high recall), they struggle with precision, resulting in reduced specificity in new sessions. This precision-recall imbalance underscores why the F1-score is the most reliable overall performance measure.

Creating a distraction detection system that works effectively for different users and VR sessions remains a challenge. Although many participants score well enough, it can be challenging to ensure consistent generalization. To address this, strategies such as transfer learning, personalized model adaptation, or identifying session-invariant distraction features may be more efficient. For real-world VR applications, it is crucial to prioritize a balanced F1 score over raw accuracy to provide both sensitivity and specificity in distraction detection.

4.3 RQ3: Participant Specific / Personalized Models

In this approach, data from the first session was used for training, while data from the second session was used for testing, and vice versa. This process was conducted across 5 participants.

The models underwent rigorous training and testing across various sequence lengths (5, 10, 15, 20, 25, 30, 35, 40, 50, 75) and multiple parameter configurations, including dropout rates (0.2, 0.3, 0.5), hidden dimension size (256,512), number of layers (4), and learning rates (0.001, 0.0001). Hyperparameter tuning was performed

to identify the optimal configuration and determine the best-performing model.

In the following results, 1 represents the first session, while 2 represents the second session. For example, "18-1" refers to the data from the 18th participant's first session, and "18-2" refers to the data from their second session. If "18-1" is seen in the results, it indicates that the model was trained on "18-2" and tested on "18-1." The following participants had two sessions, each on a different day, to collect data.

4.3.1 Participant 18

	Configuration					Accuracy		Loss		
ID	**Seq**	**DO**	**Hidden**	**Lay**	**LR**	**Train**	**Val**	**Train**	**Val**	**Test**
18-1										
C1	5	0.2	256	4	0.0001	0.999	0.999	0.227	0.227	0.229
C2	10	0.3	256	4	0.0001	0.997	0.996	0.228	0.229	0.231
C3	15	0.3	512	4	0.0001	0.996	0.994	0.229	0.230	0.229
C4	20	0.2	256	4	0.0001	0.997	0.997	0.229	0.229	0.229
C5	25	0.3	512	4	0.0001	0.993	0.983	0.230	0.234	0.224
C6	30	0.5	256	4	0.0001	0.996	0.992	0.230	0.232	0.228
C7	35	0.2	256	4	0.0001	0.990	0.990	0.231	0.231	0.225
C8	40	0.3	512	4	0.0001	1.000	0.988	0.228	0.232	0.226
C9	50	0.2	256	4	0.0001	0.982	0.964	0.234	0.237	0.244
C10	75	0.2	256	4	0.0001	1.000	0.989	0.232	0.234	0.219
18-2										
C1	5	0.3	256	4	0.0001	0.998	0.998	0.227	0.227	0.227
C2	10	0.2	256	4	0.0001	0.996	0.996	0.228	0.228	0.226
C3	15	0.3	256	4	0.0001	0.995	0.994	0.230	0.230	0.232
C4	20	0.2	256	4	0.0001	0.996	0.992	0.229	0.229	0.229
C5	25	0.3	256	4	0.0001	0.991	0.987	0.231	0.232	0.232
C6	30	0.5	256	4	0.0001	0.990	0.983	0.231	0.234	0.227
C7	35	0.5	256	4	0.0001	0.953	0.905	0.241	0.253	0.223
C8	40	0.3	256	4	0.0001	0.998	0.989	0.228	0.232	0.226
C9	50	0.5	256	4	0.0001	0.965	0.936	0.240	0.243	0.242
C10	75	0.3	512	4	0.0001	0.917	0.898	0.247	0.254	0.257

Table 9. Configuration, Accuracy, and Loss Metrics For Subject 18

For participant 18, we observe cross-session generalization effects between day 1 and day 2. When training on day 2 data and testing on day 1 (Subject 181), the best model uses seq=5 with dropout=0.2, achieving an unseen accuracy of 0.900 and F1-score of 0.947, with a notably high ROC AUC of 0.859. When reversing the training/testing order (Subject 18-2), performance becomes more consistent across

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC
18-1						
C1	0.995	0.997	0.997	0.997	0.988	0.999
C2	0.989	0.994	0.994	0.994	0.969	0.997
C3	0.992	0.996	1.000	0.991	0.996	1.000
C4	0.989	0.994	1.000	0.988	0.994	0.999
C5	1.000	1.000	1.000	1.000	1.000	1.000
C6	1.000	1.000	1.000	1.000	1.000	1.000
C7	1.000	1.000	1.000	1.000	1.000	1.000
C8	0.977	0.987	1.000	0.975	0.988	0.999
C9	0.972	0.985	0.970	1.000	0.875	0.985
C10	1.000	1.000	1.000	1.000	1.000	1.000
18-2						
C1	0.998	0.999	0.998	1.000	0.989	0.999
C2	1.000	1.000	1.000	1.000	1.000	1.000
C3	0.977	0.987	1.000	0.974	0.987	0.999
C4	0.989	0.994	1.000	0.988	0.994	0.999
C5	0.987	0.992	1.000	0.985	0.993	0.999
C6	1.000	1.000	1.000	1.000	1.000	1.000
C7	1.000	1.000	1.000	1.000	1.000	1.000
C8	0.978	0.988	1.000	0.976	0.988	0.999
C9	0.972	0.984	1.000	0.969	0.984	0.998
C10	0.917	0.950	1.000	0.905	0.952	0.994

Table 10. Test Performance Metrics For Subject 18

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC	Loss
18-1							
C1	0.900	0.947	0.900	1.000	0.859	0.985	0.288
C2	0.900	0.947	0.900	1.000	0.804	0.977	0.289
C3	0.899	0.947	0.899	1.000	0.133	0.827	0.290
C4	0.900	0.947	0.900	1.000	0.476	0.922	0.289
C5	0.898	0.946	0.898	1.000	0.775	0.971	0.290
C6	0.899	0.947	0.899	1.000	0.228	0.866	0.294
C7	0.897	0.946	0.897	1.000	0.375	0.879	0.291
C8	0.901	0.948	0.901	1.000	0.119	0.824	0.287
C9	0.899	0.000	0.000	0.000	0.211	0.860	0.398
C10	0.892	0.943	0.892	1.000	0.175	0.833	0.299
18-2							
C1	0.899	0.947	0.899	1.000	0.500	0.949	0.289
C2	0.898	0.946	0.898	1.000	0.500	0.949	0.290
C3	0.899	0.947	0.899	1.000	0.500	0.949	0.290
C4	0.898	0.946	0.898	1.000	0.500	0.949	0.291
C5	0.899	0.947	0.899	1.000	0.500	0.950	0.289
C6	0.898	0.946	0.898	1.000	0.500	0.949	0.291
C7	0.900	0.947	0.900	1.000	0.500	0.950	0.289
C8	0.900	0.947	0.900	1.000	0.500	0.950	0.289
C9	0.898	0.946	0.898	1.000	0.500	0.949	0.291
C10	0.898	0.946	0.898	1.000	0.500	0.949	0.291

Table 11. Performance Metrics on Unseen Data For Subject 18

different sequence lengths, with most configurations achieving unseen accuracy of around 0.899 and F1-scores of 0.946-0.947. However, all configurations, when trained on day 1, show an unseen ROC AUC of exactly 0.5 when tested on day 2, suggesting uniform classification confidence. As shown in Figures 7, 8, 9, 10, and

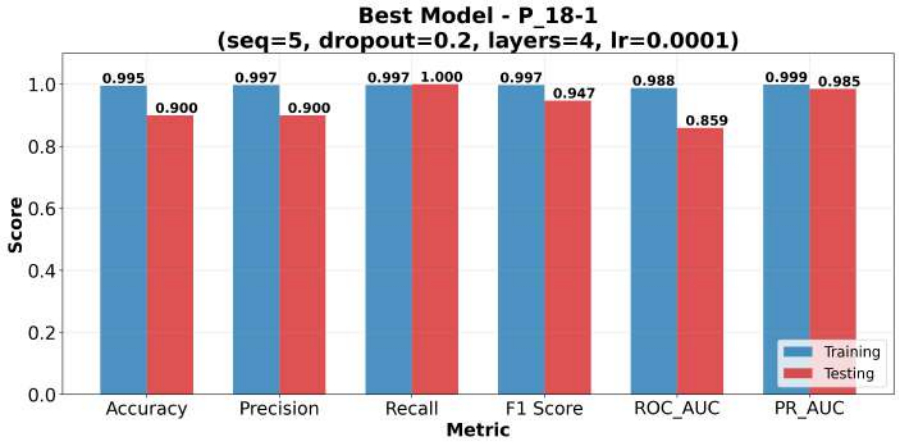


Figure 7. Train vs. Test Performance for Participant 18 on Session 1

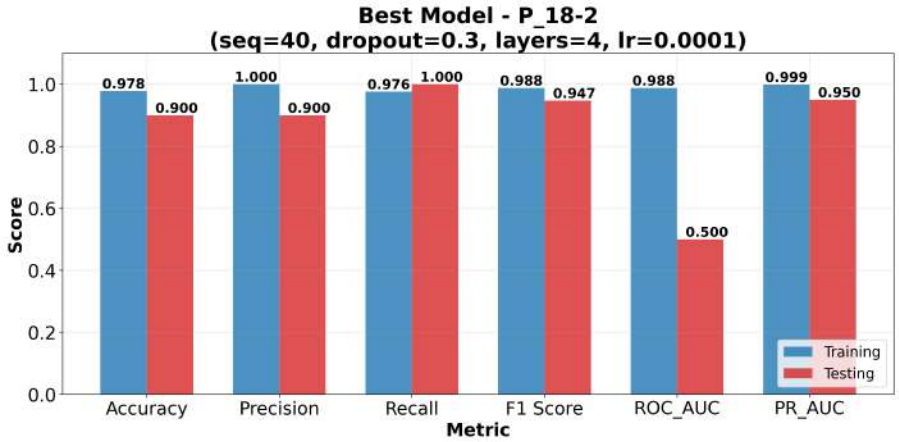


Figure 8. Train vs. Test Performance for Participant 18 on Session 2

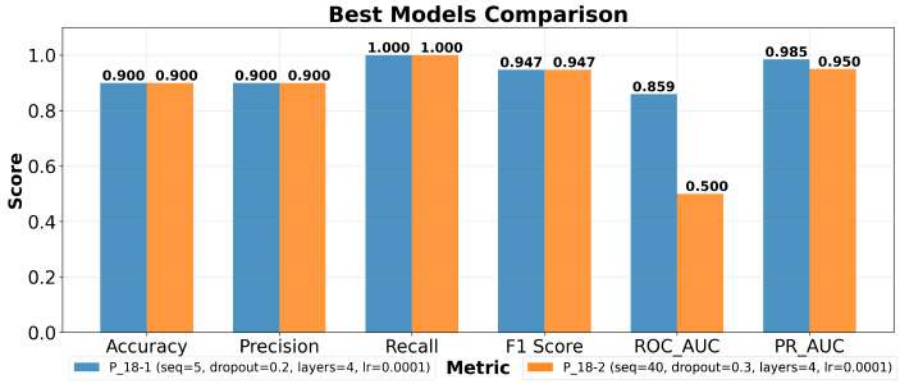


Figure 9. Cross-Data Performance Comparison for Participant 18

Tables 9, 10, 11. This asymmetry in generalization direction suggests that the day 2 data may contain more generalizable patterns, potentially due to participants'

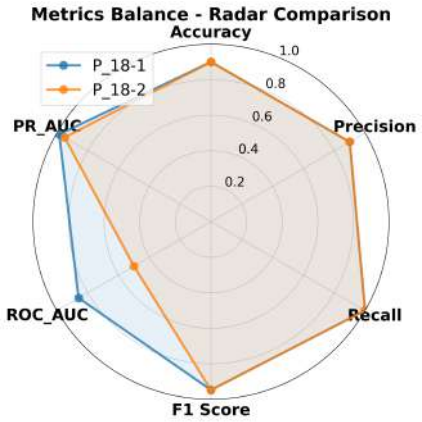


Figure 10. Multi-Metric Performance Distribution for Participant 18

familiarity with the task. This results in more stable neural signatures that better predict day 1 performance than vice versa.

4.3.2 Participant 31

	Configuration					Accuracy		Loss		
ID	**Seq**	**DO**	**Hidden**	**Lay**	**LR**	**Train**	**Val**	**Train**	**Val**	**Test**
31-1										
C1	5	0.5	512	4	0.0001	0.998	0.997	0.227	0.227	0.227
C2	10	0.5	512	4	0.0001	0.991	0.990	0.229	0.229	0.229
C3	15	0.3	256	4	0.0001	0.994	0.990	0.228	0.229	0.235
C4	20	0.5	256	4	0.0001	0.989	0.978	0.231	0.235	0.231
C5	25	0.3	256	4	0.0001	0.975	0.950	0.236	0.244	0.232
C6	30	0.5	512	4	0.0001	0.986	0.967	0.233	0.240	0.244
C7	35	0.2	256	4	0.0001	0.953	0.919	0.242	0.253	0.223
C8	40	0.5	256	4	0.0001	0.941	0.921	0.244	0.248	0.244
C9	50	0.5	256	4	0.0001	0.933	0.860	0.247	0.263	0.235
C10	75	0.2	256	4	0.0001	0.914	0.836	0.251	0.268	0.253
31-2										
C1	5	0.2	256	4	0.0001	0.998	0.996	0.366	0.366	0.366
C2	10	0.3	256	4	0.0001	0.987	0.976	0.372	0.377	0.367
C3	15	0.2	256	4	0.0001	0.994	0.984	0.368	0.372	0.368
C4	20	0.2	256	4	0.0001	0.992	0.989	0.368	0.370	0.368
C5	25	0.2	256	4	0.0001	1.000	0.983	0.365	0.372	0.374
C6	30	0.3	256	4	0.0001	0.989	0.975	0.367	0.371	0.379
C7	35	0.3	256	4	0.0001	1.000	0.995	0.363	0.366	0.382
C8	40	0.3	256	4	0.0001	0.974	0.949	0.373	0.380	0.370
C9	50	0.5	256	4	0.0001	0.921	0.859	0.394	0.418	0.405
C10	75	0.2	256	4	0.0001	0.933	0.860	0.399	0.411	0.362

Table 12. Configuration, Accuracy, and Loss Metrics For Subject 31

For Participant 31, we observe apparent asymmetry in cross-session generalization. When training on day 2 and testing on day 1 (Subject 311),

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC
31-1						
C1	0.998	0.999	0.998	1.000	0.989	0.999
C2	0.994	0.997	0.994	1.000	0.972	0.997
C3	0.961	0.978	1.000	0.957	0.978	0.998
C4	0.978	0.988	1.000	0.975	0.988	0.999
C5	0.987	0.992	1.000	0.985	0.993	0.999
C6	0.917	0.951	1.000	0.907	0.954	0.995
C7	1.000	1.000	1.000	1.000	1.000	1.000
C8	0.956	0.974	1.000	0.950	0.975	0.997
C9	1.000	1.000	1.000	1.000	1.000	1.000
C10	0.833	0.900	1.000	0.818	0.909	0.992
31-2						
C1	0.996	0.997	1.000	0.994	0.997	0.999
C2	0.994	0.996	1.000	0.992	0.996	0.999
C3	0.992	0.994	0.989	1.000	0.987	0.995
C4	0.989	0.992	1.000	0.984	0.992	0.998
C5	0.973	0.981	0.981	0.981	0.968	0.988
C6	0.967	0.976	0.976	0.976	0.960	0.985
C7	0.962	0.973	0.973	0.973	0.955	0.982
C8	1.000	1.000	1.000	1.000	1.000	1.000
C9	0.861	0.889	1.000	0.800	0.900	0.969
C10	1.000	1.000	1.000	1.000	1.000	1.000

Table 13. Test Performance Metrics For Subject 31

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC	Loss
31-1							
C1	0.708	0.769	0.682	0.881	0.350	0.594	0.525
C2	0.700	0.820	0.700	0.989	0.360	0.605	0.517
C3	0.700	0.764	0.687	0.862	0.401	0.602	0.518
C4	0.752	0.831	0.711	1.000	0.605	0.728	0.506
C5	0.707	0.825	0.702	1.000	0.425	0.633	0.514
C6	0.738	0.824	0.701	1.000	0.465	0.658	0.515
C7	0.725	0.833	0.713	1.000	0.243	0.545	0.505
C8	0.722	0.809	0.720	0.923	0.639	0.802	0.498
C9	0.702	0.051	0.125	0.032	0.241	0.549	0.562
C10	0.708	0.829	0.708	1.000	0.331	0.594	0.506
31-2							
C1	0.900	0.947	0.900	1.000	0.321	0.893	0.333
C2	0.900	0.200	1.000	0.111	0.460	0.918	0.488
C3	0.900	0.127	1.000	0.068	0.795	0.977	0.496
C4	0.900	0.747	0.856	0.663	0.316	0.891	0.419
C5	0.898	0.549	0.788	0.421	0.324	0.894	0.476
C6	0.899	0.189	1.000	0.104	0.340	0.891	0.488
C7	0.897	0.864	0.881	0.847	0.262	0.870	0.372
C8	0.897	0.946	0.897	1.000	0.323	0.889	0.336
C9	0.899	0.947	0.899	1.000	0.117	0.778	0.334
C10	0.900	0.409	0.966	0.259	0.508	0.925	0.453

Table 14. Performance Metrics on Unseen Data For Subject 31

performance is variable, with the best configuration (seq=20, dropout=0.5) achieving an unseen accuracy of 0.752 and F1-score of 0.831, along with a notable ROC AUC of 0.605. Despite lower accuracy, most configurations in this direction maintain reasonable F1-scores (0.76-0.83). When training on day 1 and testing on day 2

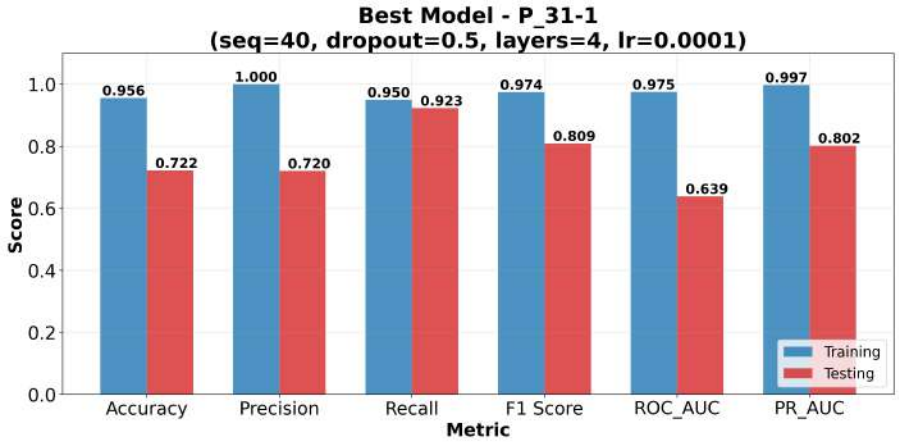


Figure 11. Train vs. Test Performance for Participant 31 on Session 1

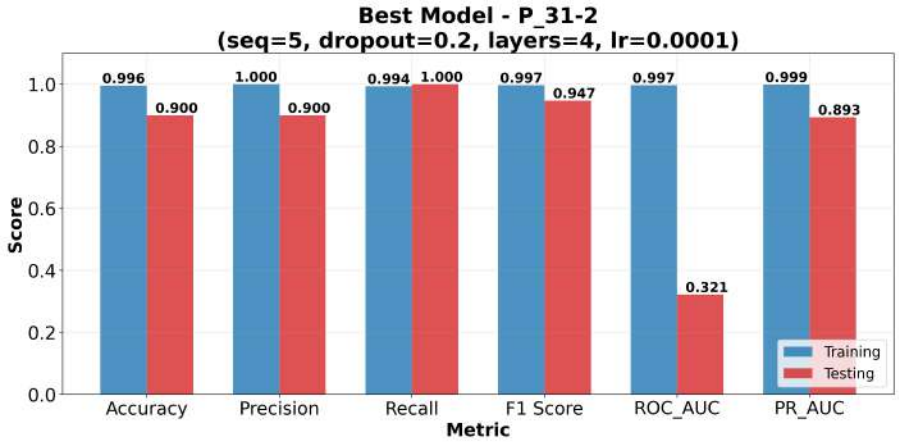


Figure 12. Train vs. Test Performance for Participant 31 on Session 2

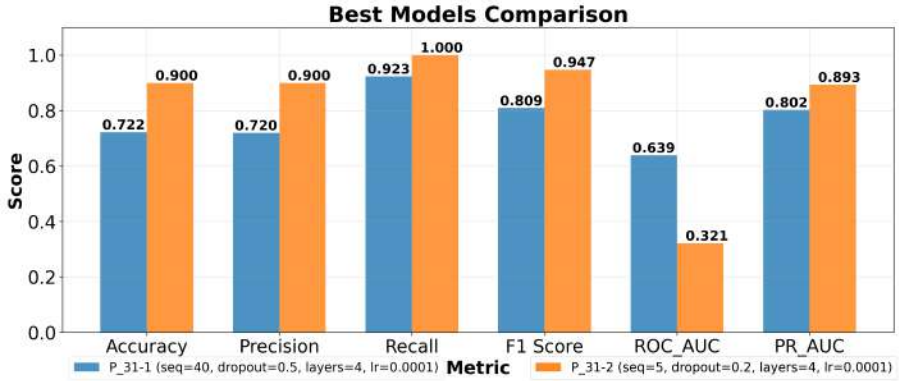


Figure 13. Cross-Data Performance Comparison for Participant 31

(Subject 312), models achieve higher unseen accuracy (around 0.90) but show extremely inconsistent F1-scores, ranging from 0.127 to 0.947. This directionality

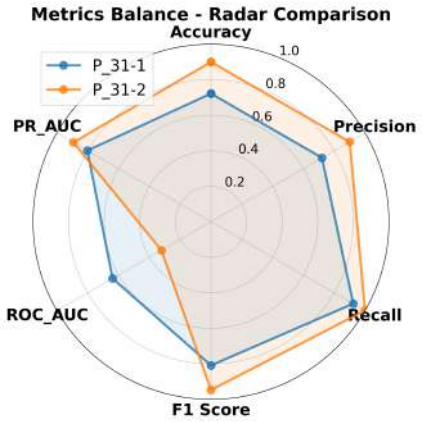


Figure 14. Multi-Metric Performance Distribution for Participant 31

effect suggests that day 1 data may contain more noise or exploratory neural patterns as the participant adjusts to the task, whereas day 2 data contain more focused, task-relevant patterns. Consequently, models trained on day 2 struggle with the noisier data from day 1, while models trained on day 1 sometimes fail to capture the more refined patterns on day 2, defaulting to majority-class predictions. As shown in Figures 11, 12, 13, 14, and Tables 12, 13, 14.

4.3.3 Participant 39

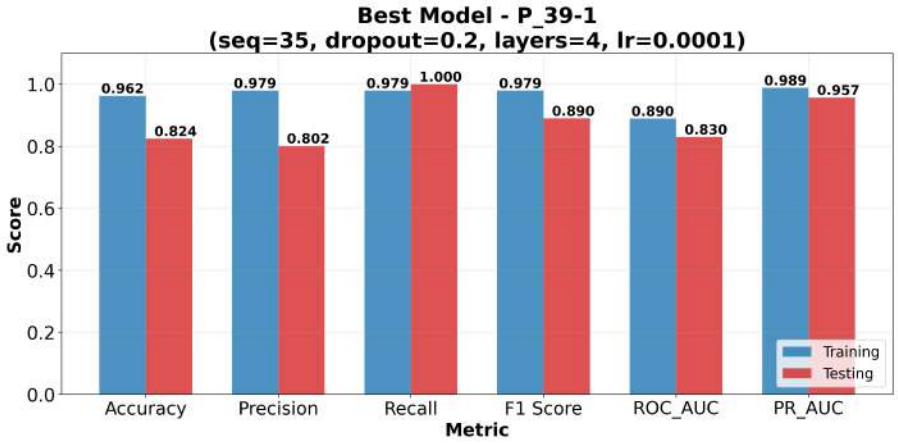


Figure 15. Train vs. Test Performance for Participant 39 on Session 1

	Configuration					Accuracy		Loss		
ID	**Seq**	**DO**	**Hidden**	**Lay**	**LR**	**Train**	**Val**	**Train**	**Val**	**Test**
39-1										
C1	5	0.3	512	4	0.0001	0.999	0.999	0.226	0.226	0.228
C2	10	0.5	256	4	0.0001	0.997	0.996	0.228	0.228	0.226
C3	15	0.2	512	4	0.0001	0.996	0.988	0.228	0.230	0.232
C4	20	0.5	256	4	0.0001	0.994	0.989	0.229	0.231	0.226
C5	25	0.5	512	4	0.0001	0.988	0.983	0.231	0.233	0.223
C6	30	0.3	256	4	0.0001	0.995	0.983	0.229	0.233	0.230
C7	35	0.2	256	4	0.0001	0.969	0.952	0.236	0.242	0.236
C8	40	0.3	256	4	0.0001	0.984	0.983	0.233	0.233	0.237
C9	50	0.3	256	4	0.0001	0.969	0.943	0.235	0.242	0.235
C10	75	0.2	256	4	0.0001	0.992	0.979	0.234	0.239	0.228
39-2										
C1	5	0.3	512	4	0.0001	0.987	0.985	0.299	0.299	0.297
C2	10	0.5	512	4	0.0001	0.981	0.976	0.300	0.302	0.298
C3	15	0.3	512	4	0.0001	0.976	0.971	0.302	0.305	0.302
C4	20	0.3	256	4	0.0001	0.988	0.981	0.300	0.304	0.309
C5	25	0.3	256	4	0.0001	0.978	0.973	0.302	0.305	0.313
C6	30	0.2	512	4	0.0001	0.967	0.954	0.305	0.309	0.303
C7	35	0.3	256	4	0.0001	0.969	0.957	0.305	0.309	0.298
C8	40	0.3	256	4	0.0001	0.978	0.944	0.307	0.315	0.310
C9	50	0.3	256	4	0.0001	0.979	0.951	0.303	0.312	0.299
C10	75	0.2	256	4	0.0001	0.981	0.959	0.302	0.309	0.311

Table 15. Configuration, Accuracy, and Loss Metrics For Subject 39

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC
39-1						
C1	0.996	0.998	0.998	0.998	0.988	0.999
C2	1.000	1.000	1.000	1.000	1.000	1.000
C3	0.977	0.987	1.000	0.974	0.987	0.999
C4	1.000	1.000	1.000	1.000	1.000	1.000
C5	1.000	1.000	1.000	1.000	1.000	1.000
C6	0.967	0.981	1.000	0.963	0.981	0.998
C7	0.962	0.979	0.979	0.979	0.890	0.989
C8	0.978	0.987	1.000	0.975	0.988	0.999
C9	1.000	1.000	1.000	1.000	1.000	1.000
C10	0.958	0.977	1.000	0.955	0.977	0.998
39-2						
C1	0.996	0.997	0.997	0.997	0.993	0.998
C2	0.989	0.993	1.000	0.986	0.993	0.999
C3	0.992	0.995	0.990	1.000	0.981	0.995
C4	0.944	0.964	1.000	0.931	0.965	0.993
C5	0.947	0.966	0.983	0.950	0.942	0.986
C6	0.967	0.979	1.000	0.958	0.979	0.996
C7	1.000	1.000	1.000	1.000	1.000	1.000
C8	0.978	0.986	0.973	1.000	0.944	0.986
C9	0.972	0.982	1.000	0.966	0.983	0.997
C10	0.958	0.973	1.000	0.947	0.974	0.995

Table 16. Test Performance Metrics For Subject 39

For Participant 39, we observe essential cross-session effects. When training on day 2 and testing on day 1 (Subject 391), performance improves with increasing sequence length, with the seq=75 configuration achieving the best-unseen accuracy

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC	Loss
39-1							
C1	0.800	0.889	0.800	1.000	0.135	0.635	0.402
C2	0.826	0.890	0.801	1.000	0.473	0.722	0.400
C3	0.800	0.826	0.781	0.877	0.662	0.901	0.421
C4	0.801	0.885	0.800	0.992	0.550	0.837	0.402
C5	0.801	0.890	0.801	1.000	0.727	0.931	0.401
C6	0.802	0.890	0.802	1.000	0.497	0.794	0.400
C7	0.824	0.890	0.802	1.000	0.830	0.957	0.400
C8	0.798	0.874	0.794	0.972	0.745	0.926	0.408
C9	0.803	0.891	0.803	1.000	0.485	0.857	0.398
C10	0.875	0.889	0.800	1.000	0.628	0.867	0.402
39-2							
C1	0.900	0.947	0.900	1.000	0.622	0.934	0.307
C2	0.900	0.947	0.900	1.000	0.518	0.919	0.308
C3	0.900	0.918	0.894	0.943	0.638	0.941	0.320
C4	0.899	0.915	0.893	0.938	0.691	0.959	0.320
C5	0.901	0.942	0.899	0.988	0.609	0.937	0.309
C6	0.899	0.772	0.913	0.668	0.599	0.937	0.349
C7	0.901	0.935	0.898	0.975	0.701	0.957	0.310
C8	0.897	0.588	0.955	0.425	0.621	0.944	0.380
C9	0.899	0.908	0.892	0.925	0.501	0.912	0.323
C10	0.900	0.185	1.000	0.102	0.580	0.940	0.440

Table 17. Performance Metrics on Unseen Data For Subject 39

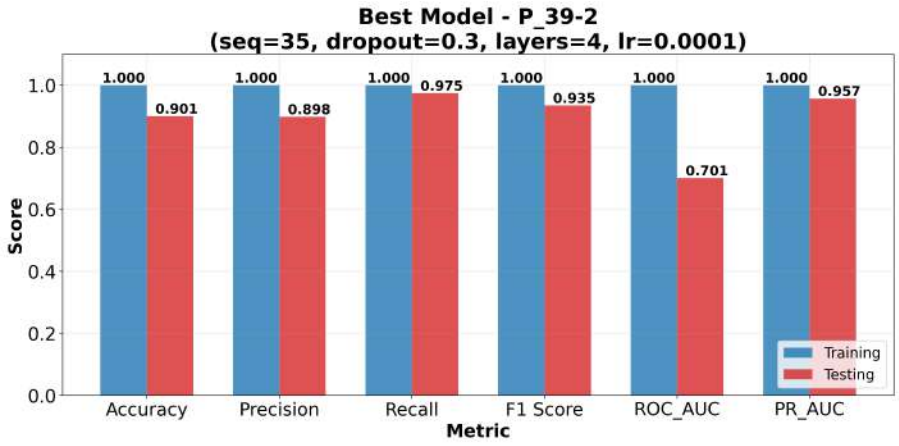


Figure 16. Train vs. Test Performance for Participant 39 on Session 2

(0.875) and F1-score (0.889). These models consistently show good recall (mostly 1.0), indicating sensitivity to positive classes in the day 1 data.

When training on day 1 and testing on day 2 (Subject 392), models maintain higher unseen accuracy (around 0.90) across configurations, but demonstrate varied F1 Scores. The strongest configuration in this direction is seq=25 with dropout=0.3, achieving an unseen accuracy of 0.901 and F1-score of 0.942. As shown in Figures

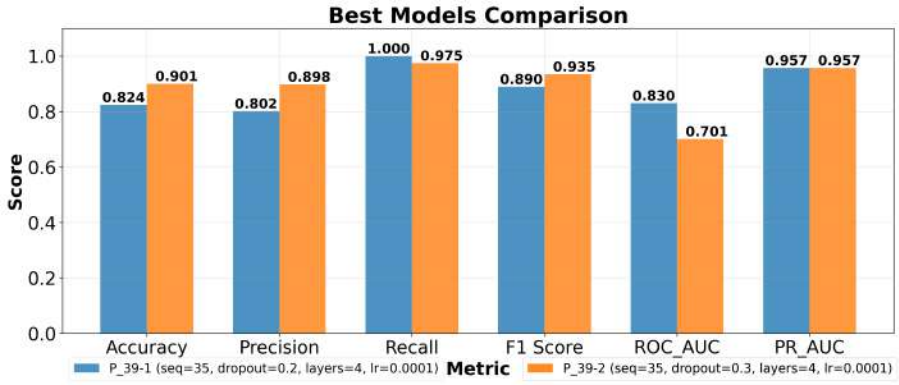


Figure 17. Cross-Data Performance Comparison for Participant 39

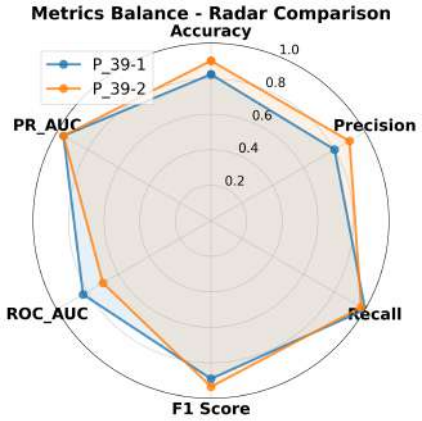


Figure 18. Multi-Metric Performance Distribution for Participant 39

15, 16, 17, 18, and Tables 15, 16, 17.

The better performance when training on day 1 suggests that this participant's day 1 neural pattern may have been more comprehensive or contained more variability, allowing better generalization to day 2. In contrast, day 2 patterns may have been more specialized and thus less generalizable to day 1.

4.3.4 Participant 40

For participant 40, we observe the strongest cross-session generalization among all participants. When training on day 2 and testing on day 1 (Subject 401),

	Configuration					Accuracy		Loss		
ID	**Seq**	**DO**	**Hidden**	**Lay**	**LR**	**Train**	**Val**	**Train**	**Val**	**Test**
40-1										
C1	5	0.2	512	4	0.0001	0.999	0.998	0.226	0.227	0.227
C2	10	0.3	256	4	0.0001	0.999	0.997	0.227	0.228	0.227
C3	15	0.3	256	4	0.0001	0.997	0.988	0.227	0.231	0.232
C4	20	0.3	256	4	0.0001	0.999	0.992	0.228	0.231	0.226
C5	25	0.3	512	4	0.0001	1.000	0.993	0.227	0.229	0.235
C6	30	0.2	256	4	0.0001	1.000	0.983	0.227	0.234	0.247
C7	35	0.5	256	4	0.0001	0.999	0.981	0.230	0.236	0.224
C8	40	0.5	512	4	0.0001	0.999	0.983	0.228	0.235	0.221
C9	50	0.5	256	4	0.001	0.955	0.950	0.242	0.249	0.232
C10	75	0.2	256	4	0.0001	0.975	0.901	0.245	0.256	0.241
40-2										
C1	5	0.5	512	4	0.001	0.999	0.999	0.226	0.226	0.226
C2	10	0.5	256	4	0.001	0.998	0.999	0.226	0.226	0.226
C3	15	0.5	256	4	0.001	0.997	0.996	0.226	0.226	0.228
C4	20	0.3	256	4	0.001	0.997	0.994	0.225	0.227	0.226
C5	25	0.5	256	4	0.001	0.995	0.987	0.226	0.230	0.225
C6	30	0.2	512	4	0.0001	1.000	0.996	0.224	0.224	0.226
C7	35	0.3	256	4	0.001	0.996	0.995	0.224	0.224	0.222
C8	40	0.3	512	4	0.0001	0.994	0.994	0.224	0.224	0.220
C9	50	0.2	512	4	0.0001	1.000	1.000	0.225	0.225	0.218
C10	75	0.5	256	4	0.001	1.000	0.989	0.222	0.228	0.216

Table 18. Configuration, Accuracy, and Loss Metrics For Subject 40

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC
40-1						
C1	0.998	0.999	0.998	1.000	0.989	0.999
C2	0.994	0.997	1.000	0.994	0.997	1.000
C3	0.992	0.996	0.991	1.000	0.962	0.996
C4	1.000	1.000	1.000	1.000	1.000	1.000
C5	0.973	0.985	1.000	0.970	0.985	0.998
C6	0.967	0.982	0.964	1.000	0.833	0.982
C7	1.000	1.000	1.000	1.000	1.000	1.000
C8	1.000	1.000	1.000	1.000	1.000	1.000
C9	1.000	1.000	1.000	1.000	1.000	1.000
C10	1.000	1.000	1.000	1.000	1.000	1.000
40-2						
C1	0.998	0.999	1.000	0.998	0.999	1.000
C2	1.000	1.000	1.000	1.000	1.000	1.000
C3	0.992	0.996	1.000	0.991	0.996	1.000
C4	1.000	1.000	1.000	1.000	1.000	1.000
C5	0.987	0.993	1.000	0.985	0.993	0.999
C6	1.000	1.000	1.000	1.000	1.000	1.000
C7	1.000	1.000	1.000	1.000	1.000	1.000
C8	1.000	1.000	1.000	1.000	1.000	1.000
C9	1.000	1.000	1.000	1.000	1.000	1.000
C10	1.000	1.000	1.000	1.000	1.000	1.000

Table 19. Test Performance Metrics For Subject 40

the models demonstrate exceptional generalization capability, with unseen accuracies consistently above 0.90 and strong F1 Scores. The standout model is seq=75 with dropout=0.2, achieving an unseen accuracy of 0.975, F1-score of 0.952,

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC	Loss
40-1							
C1	0.935	0.947	0.900	1.000	0.951	0.994	0.288
C2	0.902	0.948	0.901	1.000	0.812	0.979	0.287
C3	0.909	0.948	0.902	1.000	0.904	0.990	0.286
C4	0.904	0.896	0.900	0.891	0.734	0.965	0.300
C5	0.903	0.949	0.903	1.000	0.629	0.952	0.284
C6	0.906	0.949	0.903	1.000	0.543	0.935	0.285
C7	0.905	0.950	0.905	1.000	0.417	0.896	0.283
C8	0.942	0.951	0.906	1.000	0.934	0.993	0.282
C9	0.916	0.950	0.904	1.000	0.908	0.989	0.283
C10	0.975	0.952	0.908	1.000	0.986	0.999	0.279
40-2							
C1	0.900	0.947	0.900	1.000	0.500	0.950	0.288
C2	0.900	0.947	0.900	1.000	0.313	0.866	0.288
C3	0.900	0.947	0.900	1.000	0.508	0.912	0.288
C4	0.899	0.947	0.899	1.000	0.500	0.950	0.289
C5	0.898	0.946	0.898	1.000	0.500	0.949	0.290
C6	0.899	0.947	0.899	1.000	0.572	0.933	0.288
C7	0.897	0.946	0.897	1.000	0.419	0.931	0.291
C8	0.901	0.948	0.901	1.000	0.632	0.945	0.286
C9	0.899	0.947	0.899	1.000	0.201	0.837	0.290
C10	0.892	0.943	0.892	1.000	0.436	0.844	0.296

Table 20. Performance Metrics on Unseen Data For Subject 40

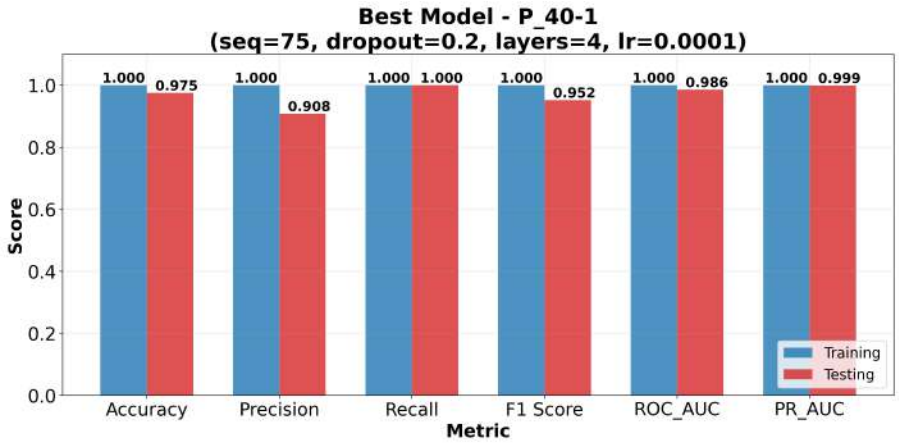


Figure 19. Train vs. Test Performance for Participant 40 on Session 1

and a remarkable ROC AUC of 0.986. A clear pattern emerges where performance improves with sequence length. As shown in Figures 19, 20, 21, 22, and Tables 18, 19, 20.

When training on day 1 and testing on day 2 (Subject 402), models achieve good unseen accuracy (around 0.90) but with significantly lower ROC AUC values (0.31-0.63). This asymmetry suggests that the day 2 neural patterns for this

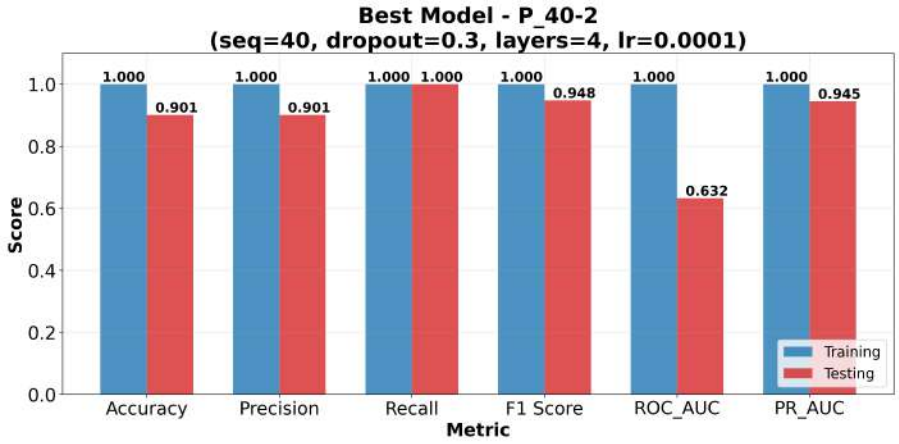


Figure 20. Train vs. Test Performance for Participant 40 on Session 2

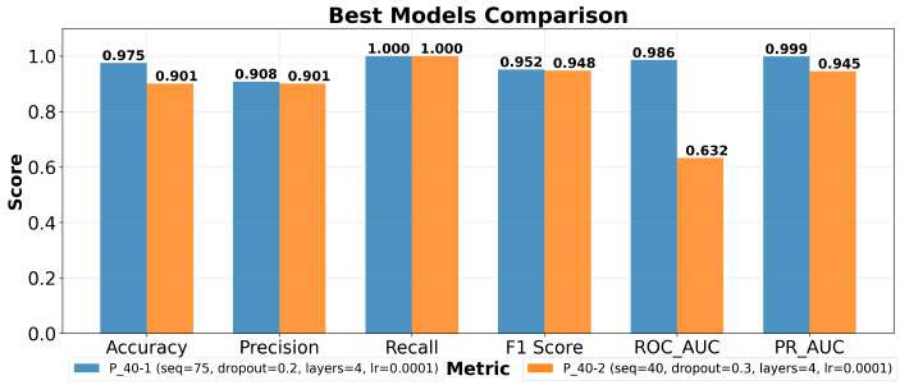


Figure 21. Cross-Data Performance Comparison for Participant 40

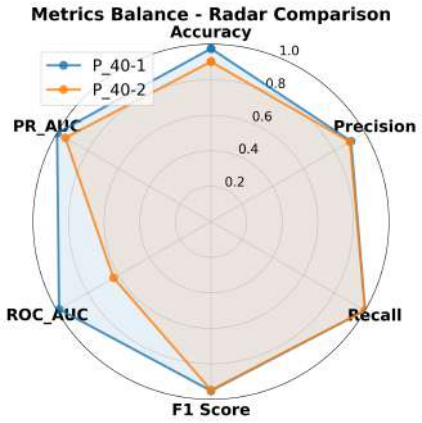


Figure 22. Multi-Metric Performance Distribution for Participant 40

participant were more stable and consistent, potentially reflecting improved familiarity with the task, which resulted in models that generalize exceptionally well to day 1

data. The strong bidirectional generalization indicates that this participant maintained relatively consistent neural patterns across sessions, making them an ideal candidate for personalized modeling approaches.

4.3.5 Participant 41

	Configuration					Accuracy		Loss		
ID	**Seq**	**DO**	**Hidden**	**Lay**	**LR**	**Train**	**Val**	**Train**	**Val**	**Test**
41-1										
C1	5	0.5	512	4	0.0001	0.989	0.987	0.296	0.297	0.297
C2	5	0.5	512	4	0.0001	0.999	0.998	0.296	0.296	0.297
C3	10	0.5	256	4	0.0001	0.991	0.980	0.298	0.302	0.299
C4	15	0.3	256	4	0.0001	0.988	0.979	0.300	0.304	0.302
C5	20	0.5	256	4	0.0001	0.983	0.966	0.301	0.306	0.303
C6	25	0.2	256	4	0.0001	0.956	0.950	0.310	0.312	0.305
C7	30	0.3	256	4	0.0001	0.965	0.942	0.305	0.315	0.305
C8	35	0.2	256	4	0.0001	0.982	0.943	0.300	0.314	0.311
C9	40	0.5	256	4	0.0001	0.961	0.904	0.311	0.328	0.297
C10	75	0.2	256	4	0.0001	0.946	0.866	0.307	0.335	0.341
41-2										
C1	5	0.3	512	4	0.0001	0.997	0.997	0.296	0.296	0.297
C2	5	0.5	256	4	0.0001	0.997	0.997	0.296	0.296	0.297
C3	10	0.3	512	4	0.0001	0.993	0.990	0.299	0.300	0.296
C4	15	0.3	512	4	0.0001	0.993	0.988	0.299	0.301	0.303
C5	20	0.5	512	4	0.0001	0.990	0.989	0.301	0.301	0.296
C6	25	0.2	512	4	0.0001	0.990	0.980	0.299	0.302	0.304
C7	30	0.2	512	4	0.0001	0.986	0.970	0.301	0.306	0.296
C8	35	0.3	256	4	0.0001	0.986	0.986	0.306	0.306	0.308
C9	40	0.5	256	4	0.0001	0.989	0.989	0.296	0.297	0.309
C10	50	0.5	256	4	0.0001	0.993	0.986	0.301	0.301	0.318
C11	75	0.3	512	4	0.0001	0.991	0.938	0.307	0.325	0.301

Table 21. Configuration, Accuracy, and Loss Metrics For Subject 41

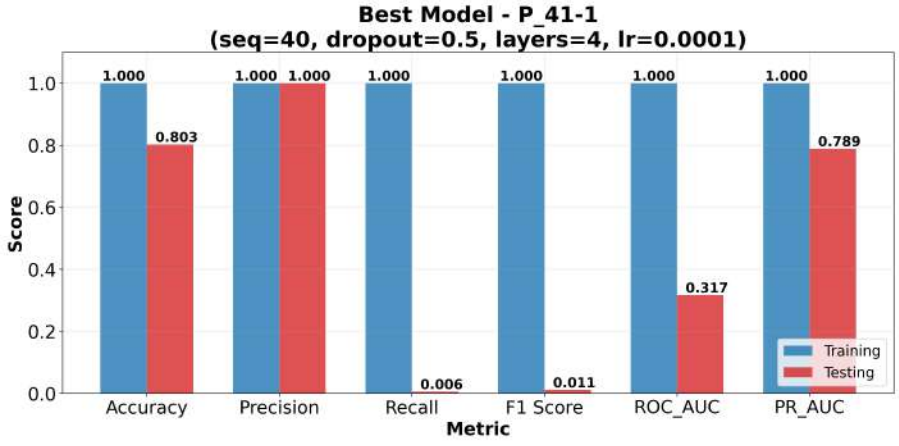


Figure 23. Train vs. Test Performance for Participant 41 on Session 1

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC
41-1						
C1	0.996	0.997	0.997	0.997	0.993	0.998
C2	0.996	0.997	0.997	0.997	0.994	0.999
C3	0.994	0.997	0.993	1.000	0.986	0.997
C4	0.992	0.995	0.990	1.000	0.981	0.995
C5	0.967	0.979	1.000	0.958	0.979	0.996
C6	0.987	0.992	0.984	1.000	0.967	0.992
C7	0.983	0.990	0.980	1.000	0.958	0.990
C8	0.906	0.938	1.000	0.884	0.942	0.989
C9	1.000	1.000	1.000	1.000	1.000	1.000
C10	0.833	0.882	1.000	0.789	0.895	0.978
41-2						
C1	0.996	0.997	0.997	0.997	0.993	0.998
C2	0.996	0.997	0.997	0.997	0.993	0.998
C3	1.000	1.000	1.000	1.000	1.000	1.000
C4	0.984	0.990	0.990	0.990	0.976	0.994
C5	1.000	1.000	1.000	1.000	1.000	1.000
C6	0.987	0.992	0.984	1.000	0.967	0.992
C7	1.000	1.000	1.000	1.000	1.000	1.000
C8	0.962	0.976	1.000	0.952	0.976	0.995
C9	0.978	0.986	0.973	1.000	0.944	0.986
C10	0.944	0.966	0.966	0.966	0.911	0.979
C11	1.000	1.000	1.000	1.000	1.000	1.000

Table 22. Test Performance Metrics For Subject 41

ID	Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC	Loss
41-1							
C1	0.800	0.000	0.000	0.000	0.192	0.671	0.485
C2	0.800	0.000	0.000	0.000	0.346	0.745	0.485
C3	0.799	0.000	0.000	0.000	0.267	0.723	0.486
C4	0.798	0.000	0.000	0.000	0.205	0.687	0.486
C5	0.799	0.000	0.000	0.000	0.492	0.783	0.486
C6	0.799	0.000	0.000	0.000	0.398	0.743	0.487
C7	0.799	0.000	0.000	0.000	0.313	0.778	0.487
C8	0.795	0.000	0.000	0.000	0.539	0.835	0.485
C9	0.803	0.011	1.000	0.006	0.317	0.789	0.485
C10	0.792	0.000	0.000	0.000	0.390	0.807	0.489
41-2							
C1	0.800	0.000	0.000	0.000	0.171	0.664	0.485
C2	0.800	0.000	0.000	0.000	0.171	0.651	0.485
C3	0.800	0.000	0.000	0.000	0.192	0.653	0.486
C4	0.799	0.000	0.000	0.000	0.160	0.671	0.485
C5	0.831	0.000	0.000	0.000	0.783	0.942	0.486
C6	0.797	0.000	0.000	0.000	0.595	0.843	0.486
C7	0.799	0.000	0.000	0.000	0.127	0.648	0.487
C8	0.802	0.000	0.000	0.000	0.160	0.647	0.485
C9	0.795	0.000	0.000	0.000	0.232	0.669	0.486
C10	0.816	0.000	0.000	0.000	0.726	0.916	0.486
C11	0.800	0.000	0.000	0.000	0.542	0.842	0.488

Table 23. Performance Metrics on Unseen Data For Subject 41

We observe the poorest cross-session generalization among all participants for participant 41. When training on day 2 and testing on day 1 (Subject 411), models achieve reasonable unseen accuracy (around 0.80) but show F1-scores of 0.000 for

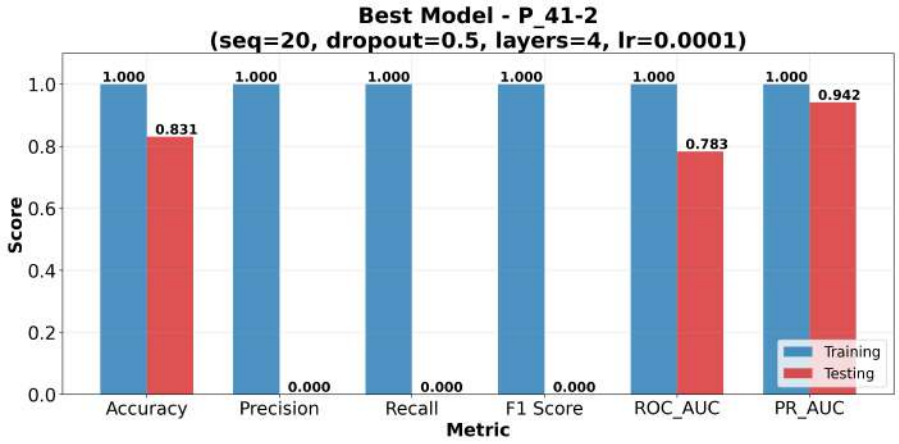


Figure 24. Train vs. Test Performance for Participant 41 on Session 2

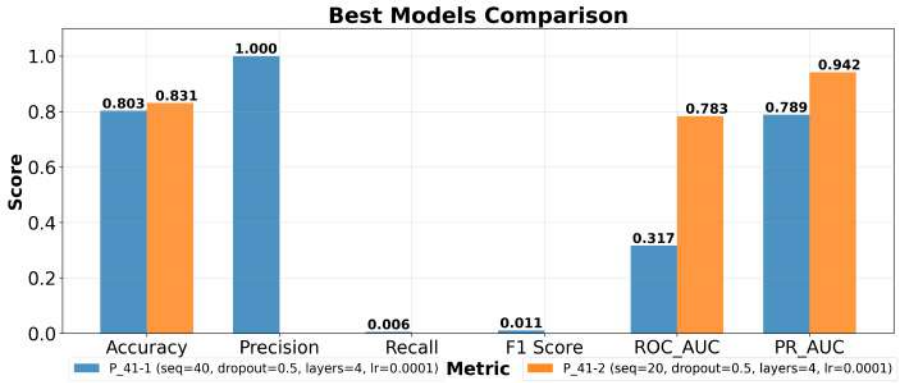


Figure 25. Cross-Data Performance Comparison for Participant 41

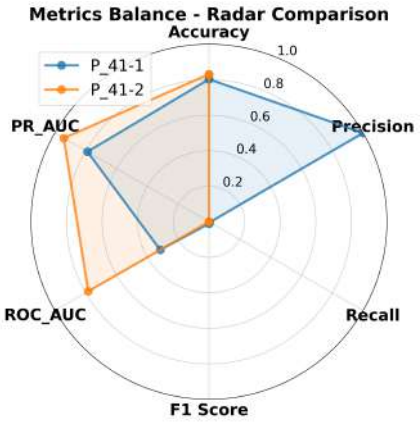


Figure 26. Multi-Metric Performance Distribution for Participant 41

almost all configurations, indicating a complete failure to maintain precision-recall balance.

Similarly, when training on day 1 and testing on day 2 (Subject 412), the only notable difference is that the seq=20 configuration achieves a higher unseen accuracy (0.8307) and ROC AUC (0.7834), but still with an F1-score of 0.000. The only configuration showing any F1-score above zero is the day 2 to day 1 seq=40 model with dropout=0.5, but even this is minimal at 0.011. As shown in Figures 23, 24, 25, 26 and Tables 21, 22, 23.

These poor bidirectional results suggest that this participant's neural patterns varied significantly between sessions, possibly due to changes in cognitive strategy, fatigue, or attention levels. This participant would likely benefit least from personalized modeling approaches that assume cross-session stability and may require more sophisticated adaptation techniques or same-day training and application.

The analysis of cross-session generalization for distraction detection reveals significant insights for developing a personalized cognitive monitoring system. Most notably, the remarkable variation in performance between individuals from Participant 40's exceptional cross-session F1-scores exceeding 0.95 to Participant 41's complete generalization failure underscores that neural and physiological responses to distraction are highly individualized. This variability manifests in asymmetrical generalization patterns between sessions, with models trained on day 2 data often showing different transfer capabilities than those trained on day 1, suggesting adaptation to the VR environment and evolving distraction response patterns. Longer sequence lengths consistently outperform shorter ones for most participants, indicating that capturing extended temporal patterns across EEG, VR interaction, and

eye movement data is crucial for building transferable models of distraction detection. The comprehensive metrics reveal cases where models achieve reasonable accuracy while failing completely in terms of precision-recall balance, particularly important in distraction detection, where both false alarms and missed distractions can significantly impact user experience. These findings collectively suggest that effective personalized distraction monitoring in VR environments must train on individual multimodal data and adapt hyperparameters to each user's specific response consistency patterns, potentially with ongoing recalibration for users demonstrating more significant inter-session variability in their distraction signatures.

The personalized modeling approach demonstrated significant variation across the five participants studied.

Three participants (18, 39, and 40) showed promising results:

- Participant 40 achieved exceptional cross-session generalization (F1-scores up to 0.952)
- Participant 18 maintained consistent F1-scores around 0.947 in both training directions
- Participant 39 achieved F1-scores up to 0.942 when training on day 1 data

Two participants showed challenges in generalization:

- Participant 31 showed asymmetrical performance (F1-scores ranging from 0.127 to 0.947 depending on training direction)

- Participant 41 demonstrated complete failure in generalization (F1-scores of 0.000 for almost all configurations)

5 Discussion and Conclusion

This chapter discusses and summarizes the contributions, limitations, and potential areas for further improvement of the present research work.

In this thesis, we developed customized models for distraction detection that do not rely on generic methods by utilizing multimodal data, including eye gaze signals, VR movement, and EEG. We experimented with cross-session validation to assess the practicality of our user-specific detection frameworks, which account for individual variations in distraction reactions.

We investigated the responses to the three research questions we put out. The experiment and its findings led us to the following answers:

1. **RQ1 (Generalized model):** Our generalized model demonstrated strong performance metrics with an accuracy of 0.930, precision of 0.929, recall of 0.974, and F1 score of 0.951 with a sequence length of 5. Despite promising results with this approach, our participant-specific model's analysis reveals its limitations.

2. **RQ2 (Participant Independent Models):** According to the participant-independent approach, Cross-session generalization varies significantly (from F1-scores $\geq$0.95 to total failure F1 = 0), indicating that distraction reactions are specific to each individual. Even while 63% of models performed rather well (F1 $\geq$ 0.6), less than half of them could maintain good detection capability (F1 $\geq$ 0.7) on unseen data. This underscores the difficulties in developing a one-size-fits-all solution.

3. **RQ3 (Personalized Models):** Longer sequence lengths that capture longer temporal patterns in multimodal data consistently resulted in better performance for most participants using our proposed personalized modeling approach. The asymmetrical generalization between day 1 and day 2 highlights the need for adaptive techniques, suggesting that distraction signatures change as people become accustomed to the VR environment.

Our approach provides a more reliable framework for practical applications in educational VR settings, as it emphasizes individual cognitive profiles. Adaptive monitoring systems that react to user-specific patterns of distraction can be developed using this method. The striking individual differences in performance underscore the highly individualized nature of neurological and physiological reactions to distraction, making it challenging to model accurately using one-size-fits-all methods.

The main contributions of this work are:

1. The crucial difference between cross-validation and practical generalization in distraction detection is shown. This was demonstrated by the variation between wildly different unseen data F1-scores (0.042 to 0.937) and consistently high cross-validation F1-scores (0.94-0.96), suggesting that conventional validation methods may provide too optimistic estimates of real-world system reliability.

2. Findings of asymmetrical training effectiveness suggest that changing distraction signatures occur as users become accustomed to the VR environment. According to our findings, there were noticeable variations in the

model's performance between training on day 1 and day 2 data, with certain subjects demonstrating superior generalization in one direction.

3. Validation that temporal sequence length significantly influences detection accuracy for personalized models. More thorough patterns in the multimodal data were consistently recorded by longer sequence lengths across participants, improving cross-session generalization for most users.

Limitations

1. The current approach requires significant initial calibration data for each user, which limits its immediate deployment for new users. This is particularly challenging for participants like Participant 41, whose neural patterns varied significantly between sessions.

2. The difference between accuracy and F1-scores demonstrates difficulties in finding a balance between precision and recall. In several instances, models obtained reasonable accuracy. Still, they failed miserably in terms of precision-recall balance (F1 score = 0), which is crucial for distraction detection, as both missed distractions and false alarms can significantly impact the user experience.

3. Personalized models' long-term stability over a long period has not been confirmed. Since our study only examined two sessions per person, it is unclear how distraction markers would change across longer periods, more VR exposures, and other VR environments.

4. The current training methodology may contain data leakage between training and validation splits, which is the reason for the inflation of validation metrics. However, since the test dataset consists of entirely unseen samples, the final performance evaluation on unseen data should remain unbiased and accurately reflect the model's generalization capabilities.

Future Work

Future studies should investigate hybrid strategies that combine personalized calibration with general baselines. For users exhibiting higher inter-session variability in their distraction signatures, exploring dynamic sequence lengths based on individual response patterns and incorporating real-time recalibration to enhance adaptability would be an interesting approach.

Additionally, exploring transfer learning strategies could reduce the initial calibration effort while preserving the advantages of personalization. Examining the neurophysiological causes of why some participants' distraction signatures remain constant while others differ significantly may help improve detection strategies.

Personalized distraction detection systems can enhance dependability by taking into account the inherent range of cognitive responses. Our work is a step towards more efficient VR-based student monitoring solutions in cognitive evaluation, adaptive learning systems, and educational environments.

Bibliography

[1] Agrawal, “Defining Immersion: Literature Review and Implications for Research on Immersive Audiovisual Experiences”. 2020.

[2] K. ALHO, D. L. WOODS, and A. ALGAZI, “Processing of auditory stimuli during auditory and visual attention as revealed by event-related potentials,” *Psychophysiology* 31.5 (1994), pp. 469–479. DOI: https://doi.org/10.1111/j.1469-8986.1994.tb01050.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1994.tb01050.x.

[3] Anubhav, D. Nath, M. Singh, D. Sethia, D. Kalra, and S. Indu, “An Efficient Approach to EEG-Based Emotion Recognition using LSTM Network”. *2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*. 2020, pp. 88–92. DOI: 10.1109/CSPA48992.2020.9068691.

[4] S. M. Asish, A. K. Kulshreshth, C. W. Borst, and S. Sutradhar, “Classification of Internal and External Distractions in an Educational VR Environment Using Multimodal Features,” *IEEE Transactions on Visualization and Computer Graphics* 30.11 (2024), pp. 7332–7342. DOI: 10.1109/TVCG.2024.3456207.

[5] C. Avila-Garzon, J. Bacca-Acosta, and J. Chaves-Rodríguez, “Predictors of Engagement in Virtual Reality Storytelling Environments about Migration,” *Applied Sciences* 13.19 (2023), p. 10915. DOI: 10.3390/app131910915.

[6] D. Beck, L. Morgado, and P. O’Shea, “Educational Practices and Strategies With Immersive Learning Environments: Mapping of Reviews for Using the Metaverse,” *IEEE Transactions on Learning Technologies* 17 (2024), pp. 319–341. DOI: 10.1109/tlt.2023.3243946.

[7] D. M. Broussard, Y. Rahman, A. K. Kulshreshth, and C. W. Borst, “An Interface for Enhanced Teacher Awareness of Student Actions and Attention in a VR Classroom”. *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2021. DOI: 10.1109/vrw52623.2021.00058.

[8] Y. Choi, J. Kim, and J.-H. Hong, “Immersion Measurement in Watching Videos Using Eye-tracking Data,” *IEEE Transactions on Affective Computing* 13.4 (2022), pp. 1759–1770. DOI: 10.1109/taffc.2022.3209311.

[9] K. T. Ciesielski, J. E. Knight, R. J. Prince, R. J. Harris, and S. D. Handmaker, “Event-related potentials in cross-modal divided attention in autism,”

Neuropsychologia 33.2 (1995), pp. 225–246.
DOI: https://doi.org/10.1016/0028-3932(94)00094-6.

[10] A. Dengel, "What Is Immersive Learning?" *2022 8th International Conference of the Immersive Learning Research Network (iLRN)*. IEEE, 2022, pp. 1–5.
DOI: 10.23919/ilrn55037.2022.9815941.

[11] J. P. Dmochowski, P. Sajda, J. Dias, and L. C. Parra, "Correlated Components of Ongoing EEG Point to Emotionally Laden Attention – A Possible Marker of Engagement?," *Frontiers in Human Neuroscience* 6 (2012).
DOI: 10.3389/fnhum.2012.00112.

[12] A. J. Dontre, "The influence of technology on academic distraction: A review," *Human Behavior and Emerging Technologies* 3.3 (2021), pp. 379–390.
DOI: https://doi.org/10.1002/hbe2.229. eprint:
https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbe2.229.

[13] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification". *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 81–84.
DOI: 10.1109/ner.2013.6695876.

[14] R. Eijlers et al., "Systematic Review and Meta-analysis of Virtual Reality in Pediatrics: Effects on Pain and Anxiety," *Anesthesia & Analgesia* 129.5 (2019), pp. 1344–1353. DOI: 10.1213/ane.0000000000004165.

[15] M. Garau, D. Friedman, H. R. Widenfeld, A. Antley, A. Brogni, and M. Slater, "Temporal and Spatial Variations in Presence: Qualitative Analysis of Interviews from an Experiment on Breaks in Presence," *Presence: Teleoperators and Virtual Environments* 17.3 (2008), pp. 293–309. DOI: 10.1162/pres.17.3.293.

[16] F. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM". *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*. Vol. 2. 1999, 850–855 vol.2.
DOI: 10.1049/cp:19991218.

[17] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks". *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Y. W. Teh and M. Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, 2010, pp. 249–256.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

[19] X. Hu, J. Chen, F. Wang, and D. Zhang, "Ten challenges for EEG-based affective computing," *Brain Science Advances* 5.1 (2019), pp. 1–20. DOI: 10.1177/2096595819896200.

[20] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer Learning in Brain-Computer Interfaces," *IEEE Computational Intelligence Magazine* 11.1 (2016), pp. 20–31. DOI: 10.1109/mci.2015.2501545.

[21] A. Kashevnik, R. Shchedrin, C. Kaiser, and A. Stocker, "Driver Distraction Detection Methods: A Literature Review and Framework," *IEEE Access* 9 (2021), pp. 60063–60076. DOI: 10.1109/access.2021.3073599.

[22] S. Katsigiannis and N. Ramzan, "DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices," *IEEE Journal of Biomedical and Health Informatics* 22.1 (2018), pp. 98–107. DOI: 10.1109/jbhi.2017.2688239.

[23] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[24] S. Koelstra et al., "DEAP: A Database for Emotion Analysis; Using Physiological Signals," *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 18–31. DOI: 10.1109/t-affc.2011.15.

[25] T. Koenig, D. Studer, D. Hubl, L. Melie, and W. Strik, "Brain connectivity at different time-scales measured with EEG," *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1457 (2005), pp. 1015–1024. DOI: 10.1098/rstb.2005.1649. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2005.1649.

[26] N. Kosmyna and P. Maes, "AttentivU: An EEG-Based Closed-Loop Biofeedback System for Real-Time Monitoring and Improvement of Engagement for Personalized Learning," *Sensors* 19.23 (2019). DOI: 10.3390/s19235200.

[27] L. R. Krol, P. Haselager, and T. O. Zander, "Erratum: Cognitive and affective probing: a tutorial and review of active learning for neuroadaptive technology (2020 J. Neural Eng. 17 012001)," *Journal of Neural Engineering* 17.4 (2020), p. 049601. DOI: 10.1088/1741-2552/ab8a6f.

[28] Q. V. Le, N. Jaitly, and G. E. Hinton, *A Simple Way to Initialize Recurrent Networks of Rectified Linear Units*. 2015. arXiv: 1504.00941 [cs.NE].

[29] D. Lehmann and W. Skrandies, “Reference-free identification of components of checkerboard-evoked multichannel potential fields,” *Electroencephalography and Clinical Neurophysiology* 48.6 (1980), pp. 609–621. DOI: 10.1016/0013-4694(80)90419-8.

[30] A. Levordashka, D. Stanton Fraser, and I. D. Gilchrist, “Measuring real-time cognitive engagement in remote audiences,” *Scientific Reports* 13.1 (2023). DOI: 10.1038/s41598-023-37209-7.

[31] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, “Domain Adaptation for EEG Emotion Recognition Based on Latent Representation Similarity,” *IEEE Transactions on Cognitive and Developmental Systems* 12.2 (2020), pp. 344–353. DOI: 10.1109/tcds.2019.2949306.

[32] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, and H. He, “Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition,” *IEEE Transactions on Cybernetics* (2019), pp. 1–13. DOI: 10.1109/tcyb.2019.2904052.

[33] J. Liu, X. Shen, S. Song, and D. Zhang, “Domain Adaptation for Cross-Subject Emotion Recognition by Subject Clustering”. *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2021, pp. 904–908. DOI: 10.1109/ner49283.2021.9441368.

[34] B.-Q. Ma, H. Li, W.-L. Zheng, and B.-L. Lu, “Reducing the Subject Variability of EEG Signals with Adversarial Domain Generalization”. *Neural Information Processing*. Springer International Publishing, 2019, pp. 30–42. DOI: 10.1007/978-3-030-36708-4_3.

[35] K. M. Malloy and L. S. Milling, “The effectiveness of virtual reality distraction for pain reduction: A systematic review,” *Clinical Psychology Review* 30.8 (2010), pp. 1011–1018. DOI: 10.1016/j.cpr.2010.07.001.

[36] Y. I. Nakano and R. Ishii, “Estimating user’s engagement from eye-gaze behaviors in human-agent conversations”. *Proceedings of the 15th international conference on Intelligent user interfaces*. IUI ’10. ACM, 2010. DOI: 10.1145/1719970.1719990.

[37] S. Niknam and J. Botev, “Predicting Cognitive Failures in Virtual Reality Using Pupillometry”. *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE, 2024, pp. 261–264. DOI: 10.1109/aixvr59861.2024.00043.

[38] J. O’Hagan, J. R. Williamson, F. Mathis, M. Khamis, and M. McGill, “Re-Evaluating VR User Awareness Needs During Bystander Interactions”.

Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. ACM, 2023, pp. 1–17. DOI: 10.1145/3544548.3581018.

[39] R. Pascanu, T. Mikolov, and Y. Bengio, *On the difficulty of training Recurrent Neural Networks*. 2013. arXiv: 1211.5063 [cs.LG].

[40] M. Á. Pérez-Juárez, D. González-Ortega, and J. M. Aguiar-Pérez, "Digital Distractions from the Point of View of Higher Education Students," *Sustainability* 15.7 (2023). DOI: 10.3390/su15076044.

[41] Y. Rahman, S. M. Asish, N. P. Fisher, E. C. Bruce, A. K. Kulshreshth, and C. W. Borst, "Exploring Eye Gaze Visualization Techniques for Identifying Distracted Students in Educational VR". *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2020, pp. 868–877. DOI: 10.1109/VR46266.2020.00009.

[42] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive Learning of Subject-Invariant EEG Representations for Cross-Subject Emotion Recognition," *IEEE Transactions on Affective Computing* 14.3 (2023), pp. 2496–2511. DOI: 10.1109/taffc.2022.3164516.

[43] M. Slater, V. Linakis, M. Usoh, and R. Kooper, "Immersion, presence and performance in virtual environments: an experiment with tri-dimensional chess". *Proceedings of the ACM Symposium on Virtual Reality Software and Technology - VRST '96*. VRST '96. ACM Press, 1996, pp. 163–172. DOI: 10.1145/3304181.3304216.

[44] M. Slater and A. Steed, "A Virtual Presence Counter," *Presence: Teleoperators and Virtual Environments* 9.5 (2000), pp. 413–434. DOI: 10.1162/105474600566925.

[45] D. J. Strauss, A. L. Francis, J. Vibell, and F. I. Corona–Strauss, "The role of attention in immersion: The two–competitor model," *Brain Research Bulletin* 210 (2024), p. 110923. DOI: 10.1016/j.brainresbull.2024.110923.

[46] H. Tadayyoni, M. S. Ramirez Campos, A. J. U. Quevedo, and B. A. Murphy, "Biomarkers of Immersion in Virtual Reality Based on Features Extracted from the EEG Signals: A Machine Learning Approach," *Brain Sciences* 14.5 (2024), p. 470. DOI: 10.3390/brainsci14050470.

[47] Y. Tao and P. Lopes, "Integrating Real-World Distractions into Virtual Reality". *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. UIST '22. ACM, 2022, pp. 1–16. DOI: 10.1145/3526113.3545682.

[48] R. Venkatakrishnan, R. Venkatakrishnan, B. Raveendranath, D. M. Sarno, A. C. Robb, W.-C. Lin, and S. V. Babu, “The Effects of Auditory, Visual, and Cognitive Distractions on Cybersickness in Virtual Reality,” *IEEE Transactions on Visualization and Computer Graphics* 30.8 (2024), pp. 5350–5369. DOI: 10.1109/tvcg.2023.3293405.

[49] X. Wang et al., “Fusion of Multi-domain EEG Signatures Improves Emotion Recognition,” *JIN* 23.1 (2024), 18–null. DOI: 10.31083/j.jin2301018.

[50] B. G. Witmer and M. J. Singer, “Measuring Presence in Virtual Environments: A Presence Questionnaire,” *Presence: Teleoperators and Virtual Environments* 7.3 (1998), pp. 225–240. DOI: 10.1162/105474698565686.

[51] S. Zhang, X. Feng, and Y. Shen, “Quantifying Auditory Presence Using Electroencephalography,” *Applied Sciences* 11.21 (2021), p. 10461. DOI: 10.3390/app112110461.

[52] W.-L. Zheng and B.-L. Lu, “Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks,” *IEEE Transactions on Autonomous Mental Development* 7.3 (2015), pp. 162–175. DOI: 10.1109/tamd.2015.2431497.

[53] Y. Zhu, Q. Wang, and L. Zhang, “Study of EEG characteristics while solving scientific problems with different mental effort,” *Scientific Reports* 11.1 (2021), p. 23783. DOI: 10.1038/s41598-021-03321-9.

6 Appendices

6.1 Appendix A: Window Length Parameters

This appendix details the relationship between window size parameters and their corresponding temporal durations for the signal processing applied in this study.

Table 24. Window Size Parameters and Their Corresponding Temporal Durations

Window Size	Overlap	Duration (ms)	Total Duration (s)
5	2	55	0.055
10	5	111	0.111
15	7	167	0.167
20	10	222	0.222
25	12	278	0.278
30	15	333	0.333
35	17	389	0.389
40	20	444	0.444
50	25	556	0.556
75	37	833	0.833

6.2 Appendix B: Comparison Graphs for Research Question 1

This appendix presents comprehensive visualization of model performance metrics for the first research question. Each plot visualizes model performance against the corresponding metric, giving direct comparative analysis.

6.2.1 Model Configurations

Table 25. Model Configuration Mapping

ID	Seq Length	Dropout	Hidden Size	Layers	Learning Rate
C1	5	0.5	256	4	0.0001
C2	10	0.5	256	4	0.0001
C3	15	0.5	512	4	0.0001
C4	20	0.5	512	4	0.0001
C5	25	0.3	256	4	0.0001
C6	30	0.5	256	4	0.0001
C7	35	0.5	256	4	0.0001
C8	40	0.5	256	4	0.0001
C9	50	0.2	256	4	0.0001
C10	75	0.5	256	4	0.0001

6.2.2 Performance Metrics Across Models

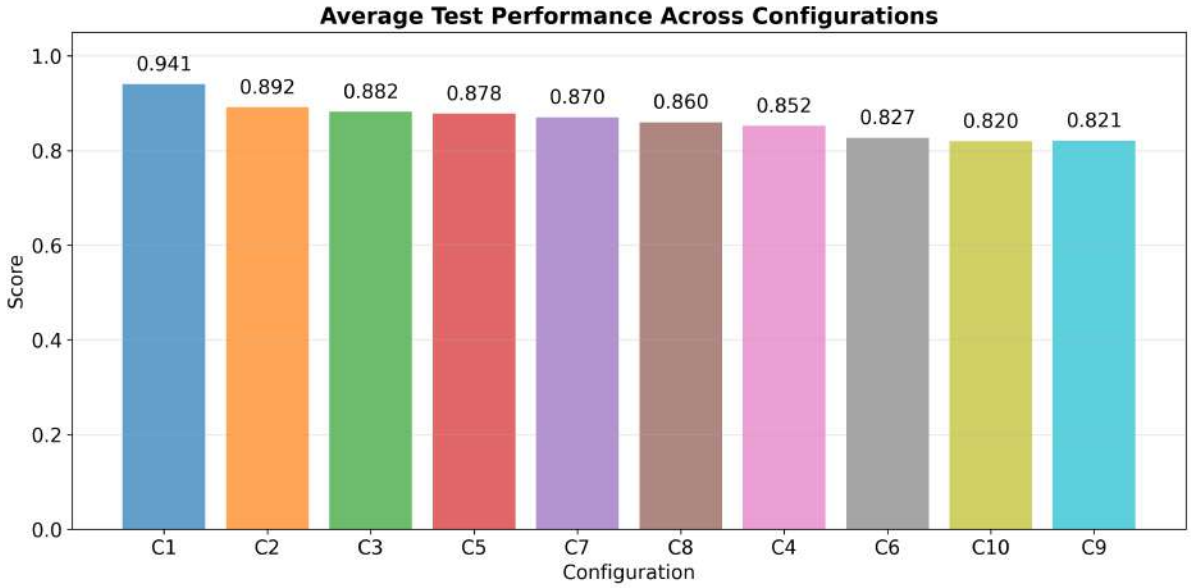


Figure 27. Average Test Performance

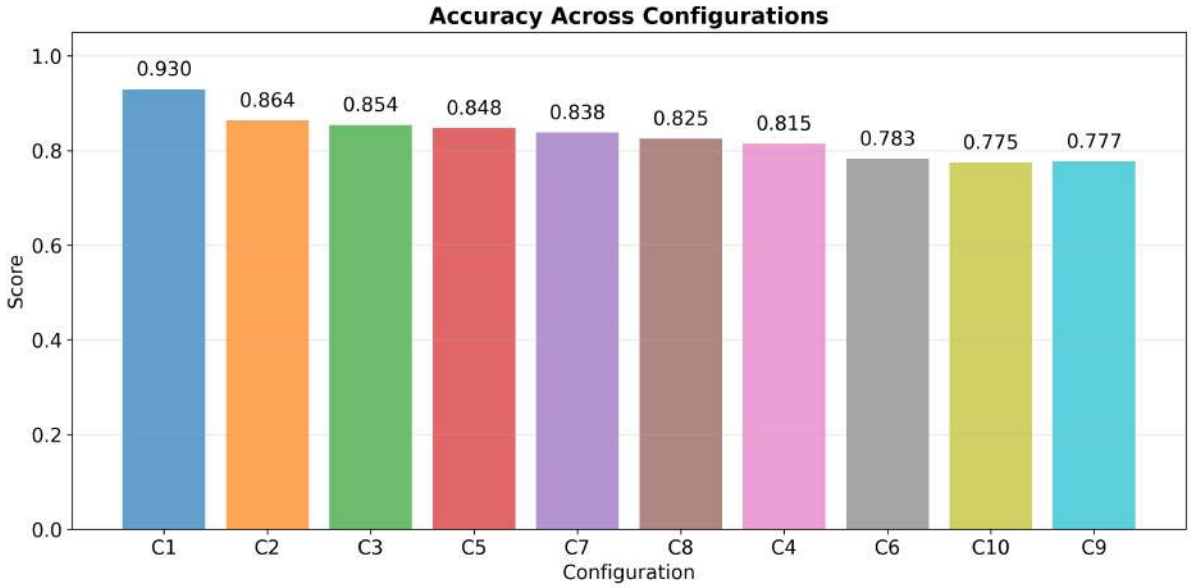


Figure 28. Test Accuracy

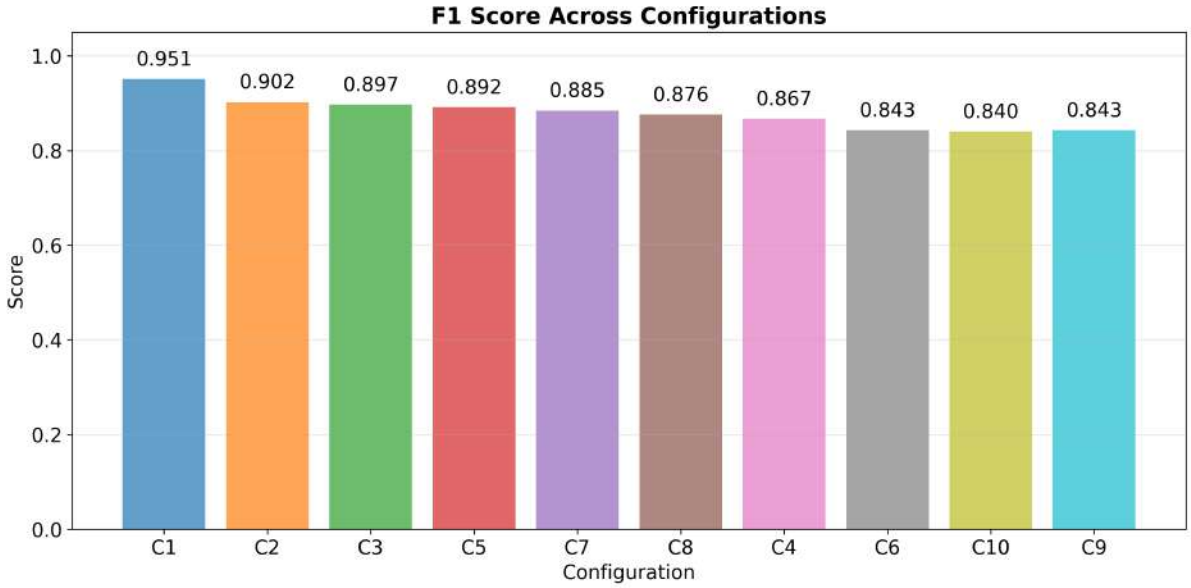


Figure 29. Test F1 Score

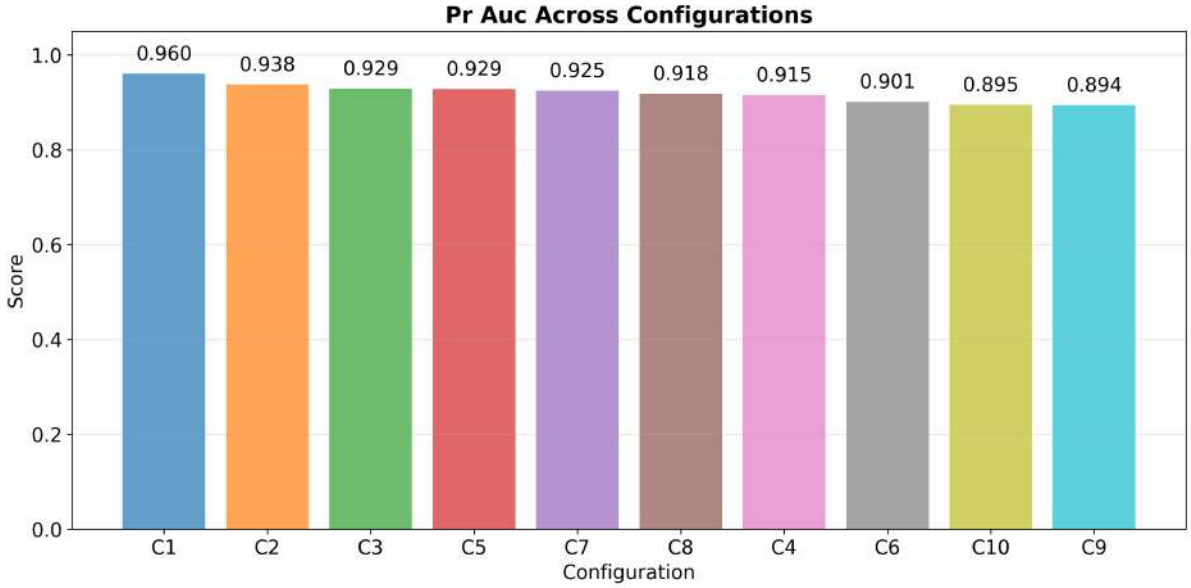


Figure 30. Test PR AUC

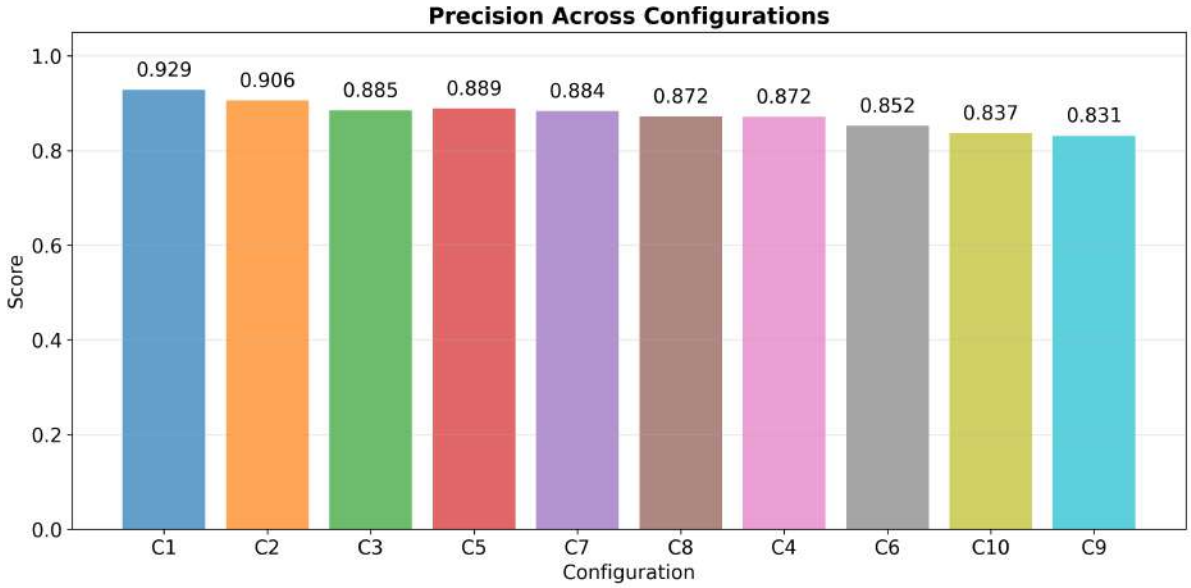


Figure 31. Test Precision

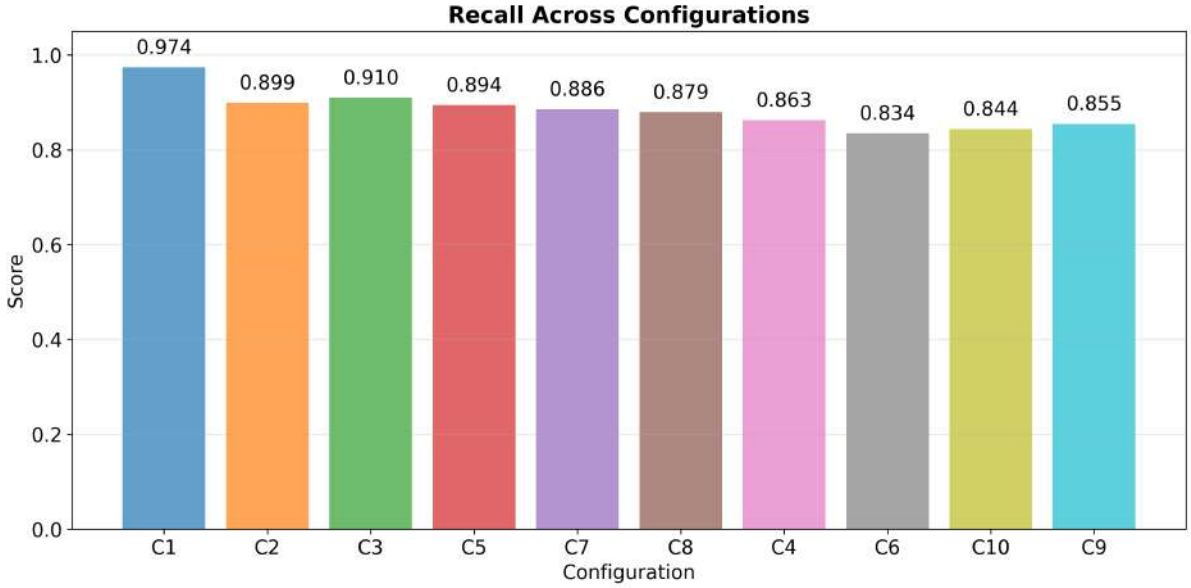


Figure 32. Test Recall

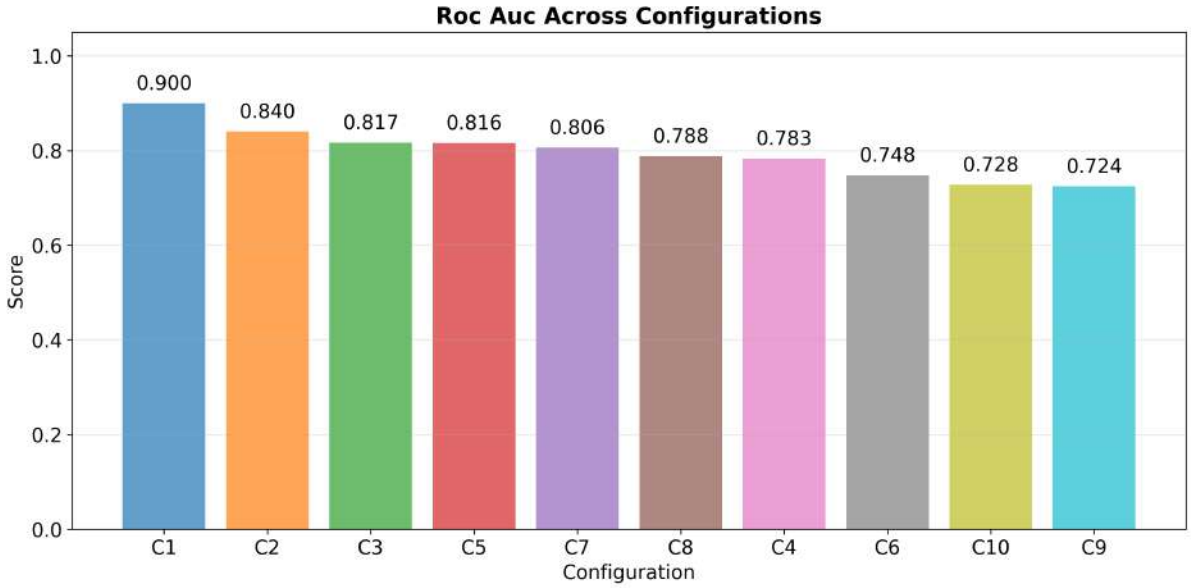


Figure 33. Test ROC AUC

6.2.3 ROC Curve Analysis

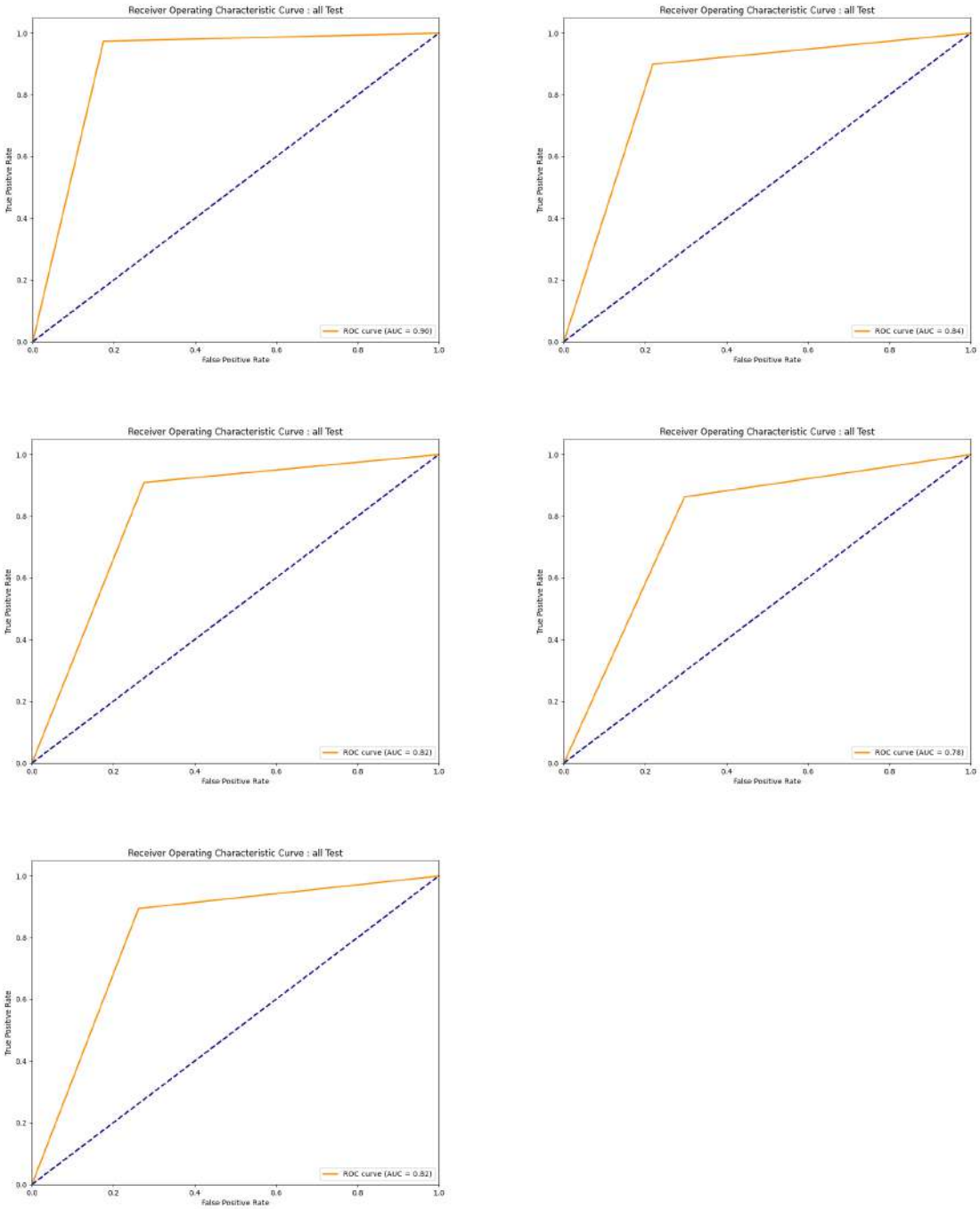


Figure 34. ROC curves for generalized models C1-C5 on test data. *Row 1:* C1 (left) and C2 (right). *Row 2:* C3 (left) and C4 (right). *Row 3:* C5.

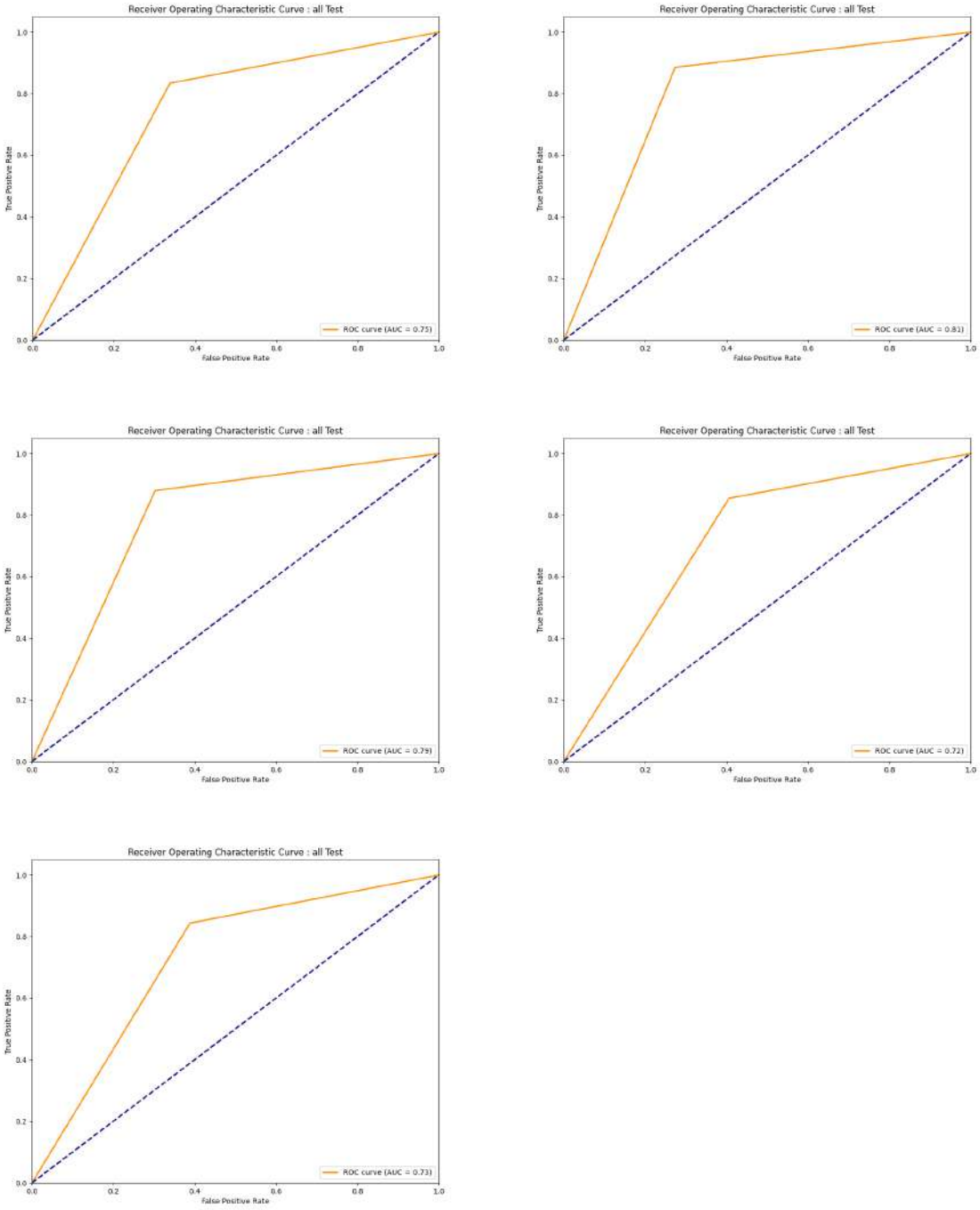


Figure 35. ROC curves for generalized models C6-C10 on test data. *Row 1:* C6 (left) and C7 (right). *Row 2:* C8 (left) and C9 (right). *Row 3:* C10.

6.2.4 Precision-Recall Analysis

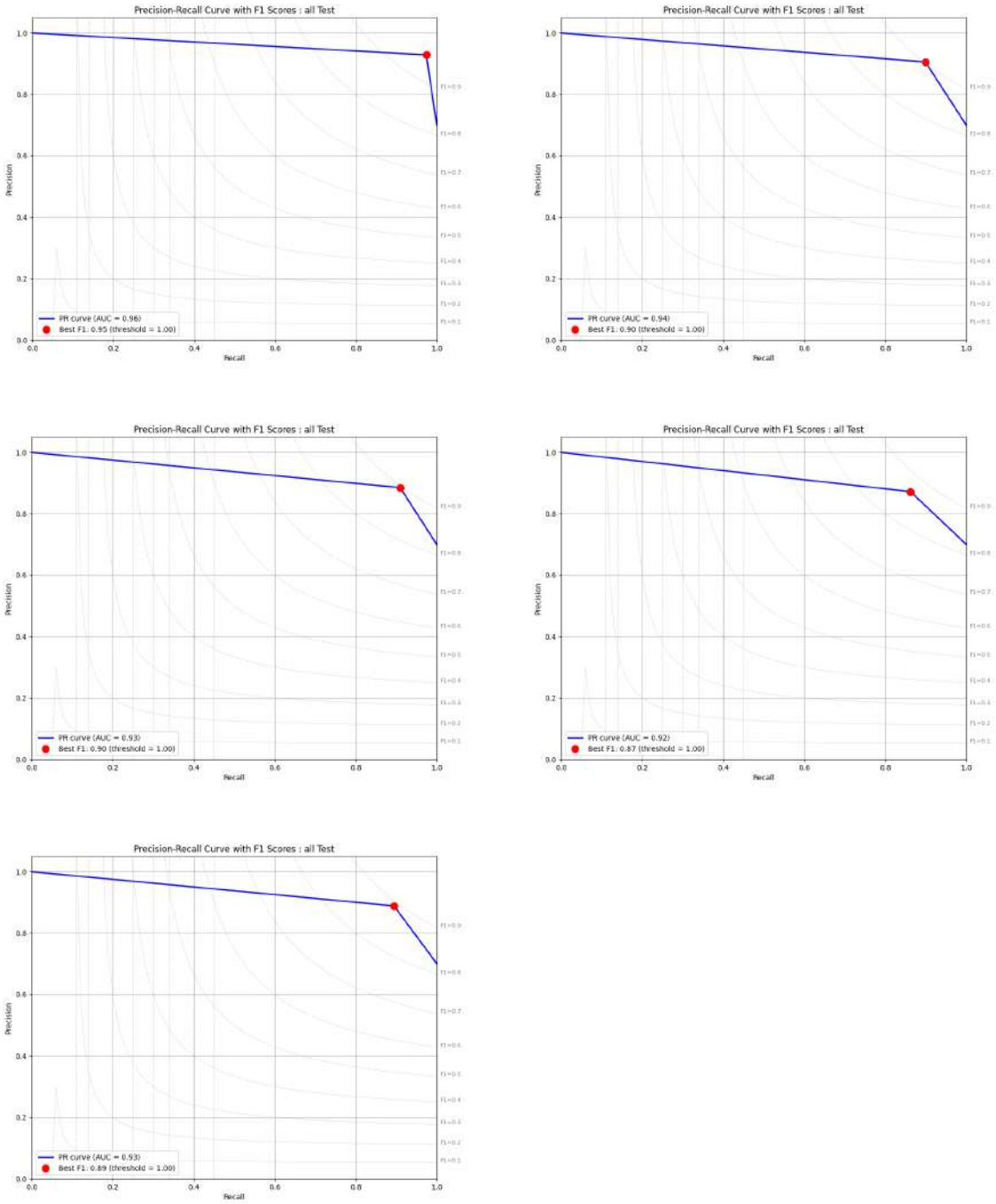


Figure 36. Precision-Recall curves for generalized models C1-C5 on test data. *Row 1:* C1 (left) and C2 (right). *Row 2:* C3 (left) and C4 (right). *Row 3:* C5.

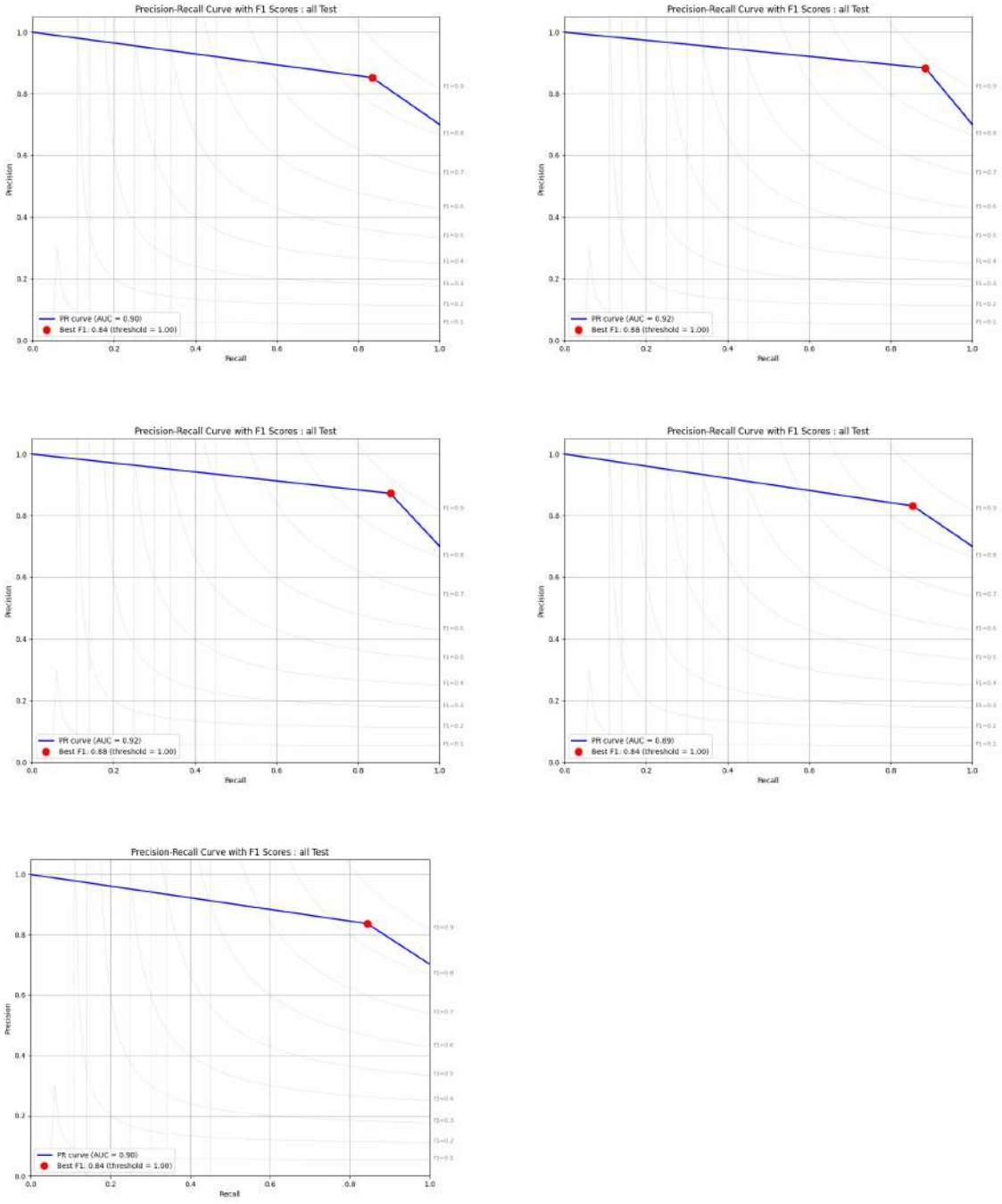


Figure 37. Precision-Recall curves for generalized models C6-C10 on test data. *Row 1:* C6 (left) and C7 (right). *Row 2:* C8 (left) and C9 (right). *Row 3:* C10.

6.2.5 Confusion Matrix Evaluation

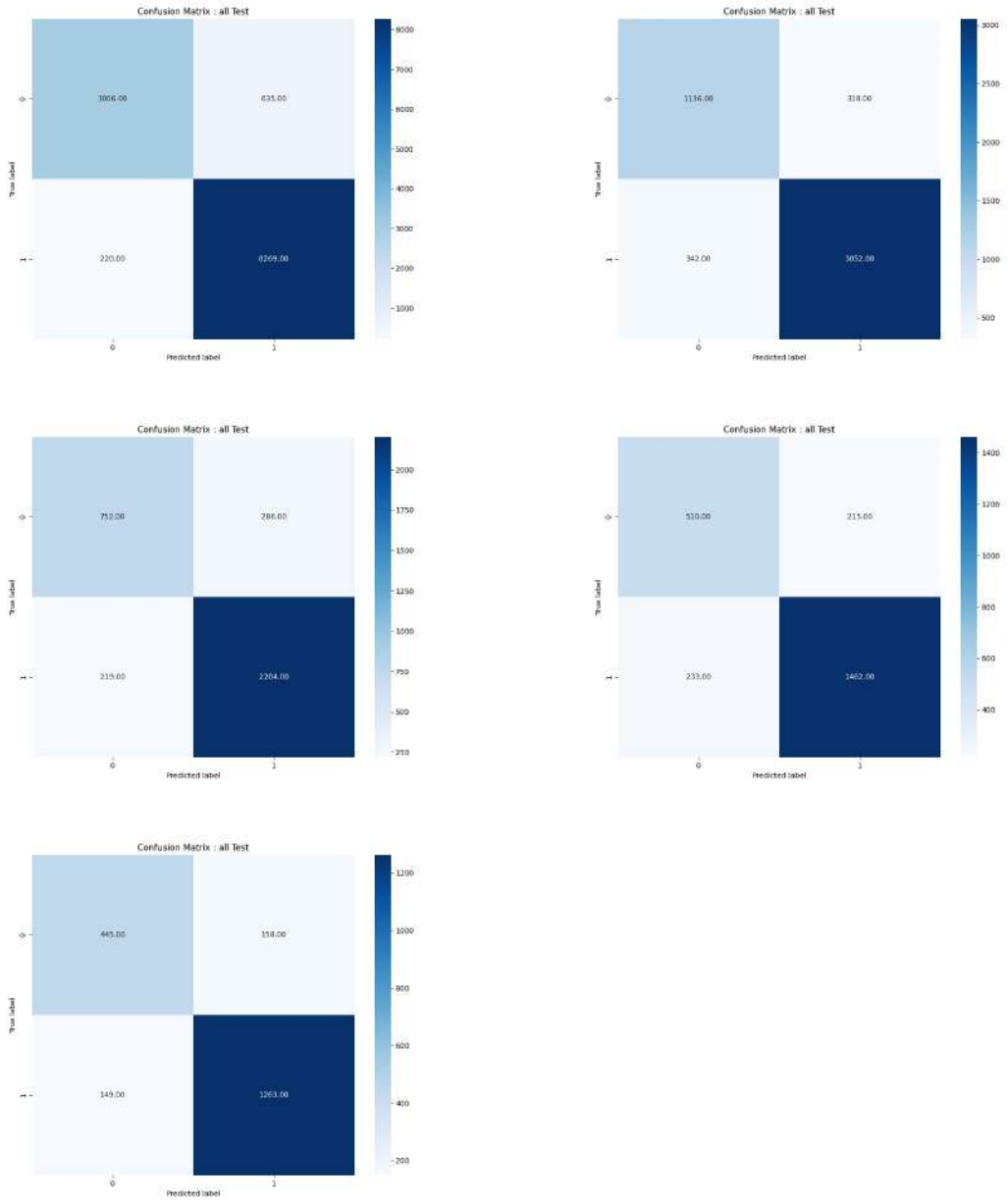


Figure 38. Confusion matrices for generalized models C1-C5 on test data. *Row 1:* Models C1 (left) and C2 (right). *Row 2:* Models C3 (left) and C4 (right). *Row 3:* Model C5.

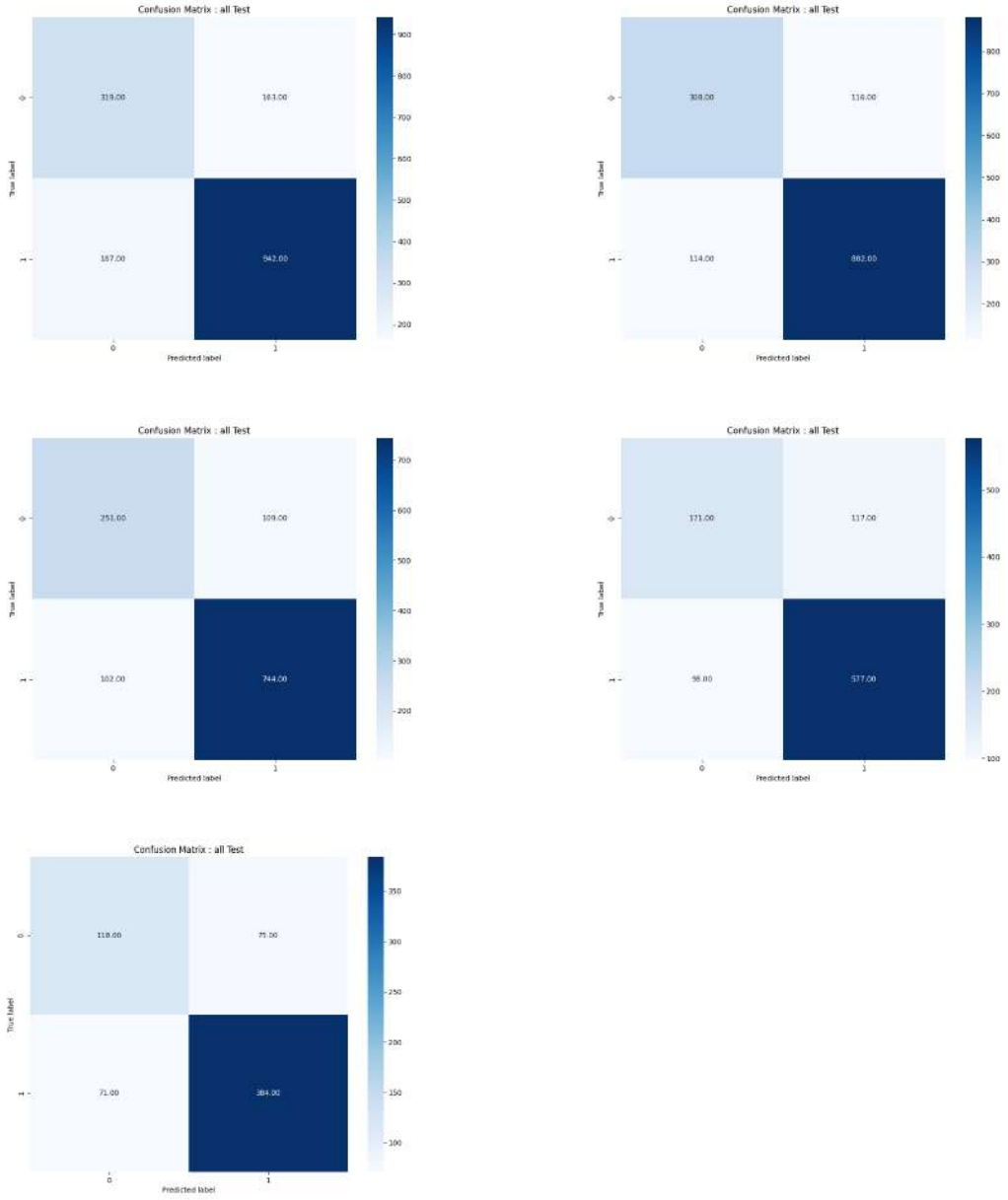


Figure 39. Confusion matrices for generalized models C6-C10 on test data. *Row 1:* C6 (left) and C7 (right). *Row 2:* C8 (left) and C9 (right). *Row 3:* C10.

ok

6.3 Appendix C: Graphs for Research Question 2

6.3.1 Accuracy and F1-Score Metrics

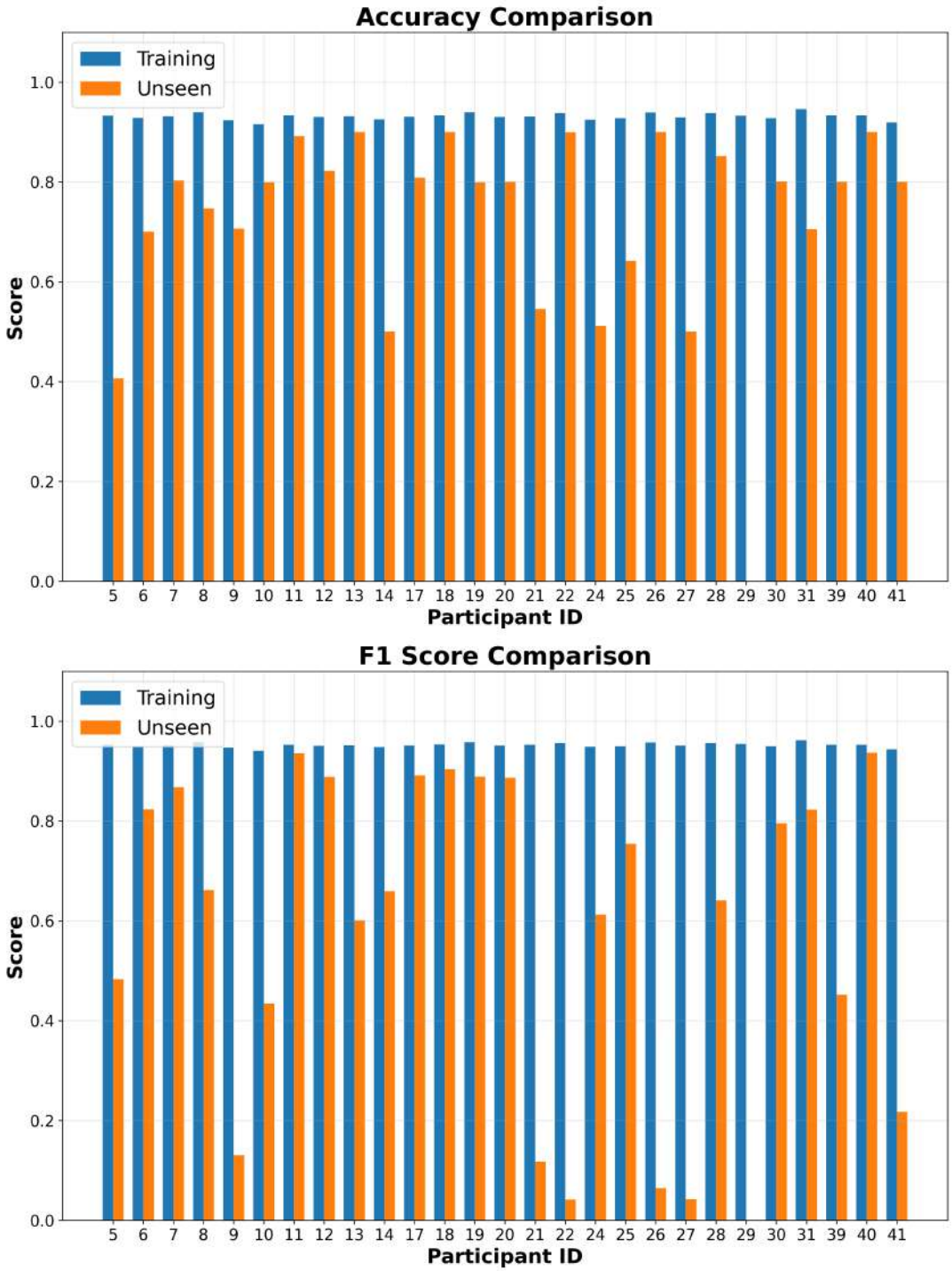


Figure 40. Primary performance metrics. *Top:* Accuracy *Bottom:* F1 Score

6.3.2 ROC AUC and PR AUC Analysis

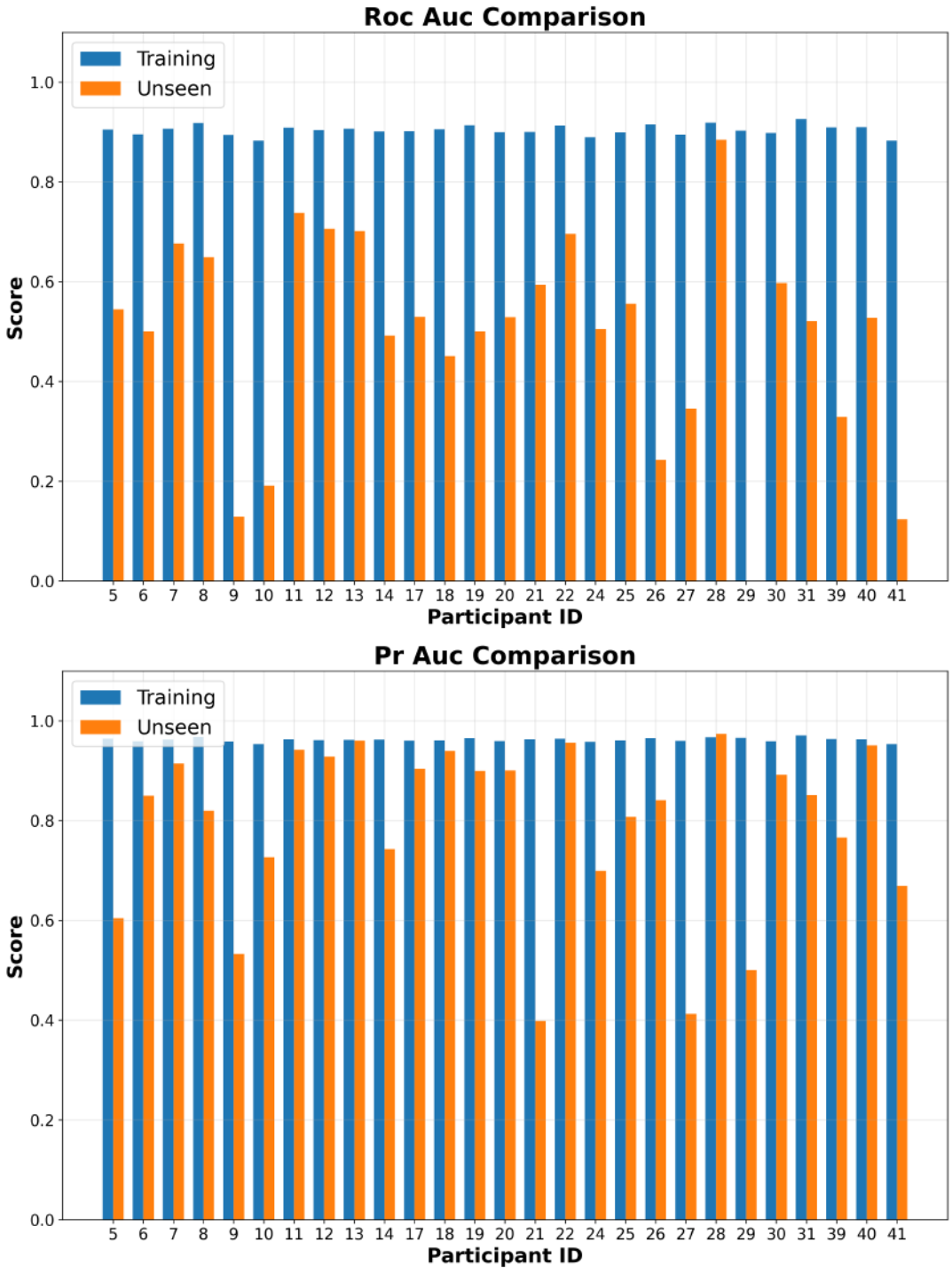


Figure 41. Area Under Curve metrics. *Top:* ROC-AUC. *Bottom:* PR-AUC.

6.3.3 Precision and Recall Analysis

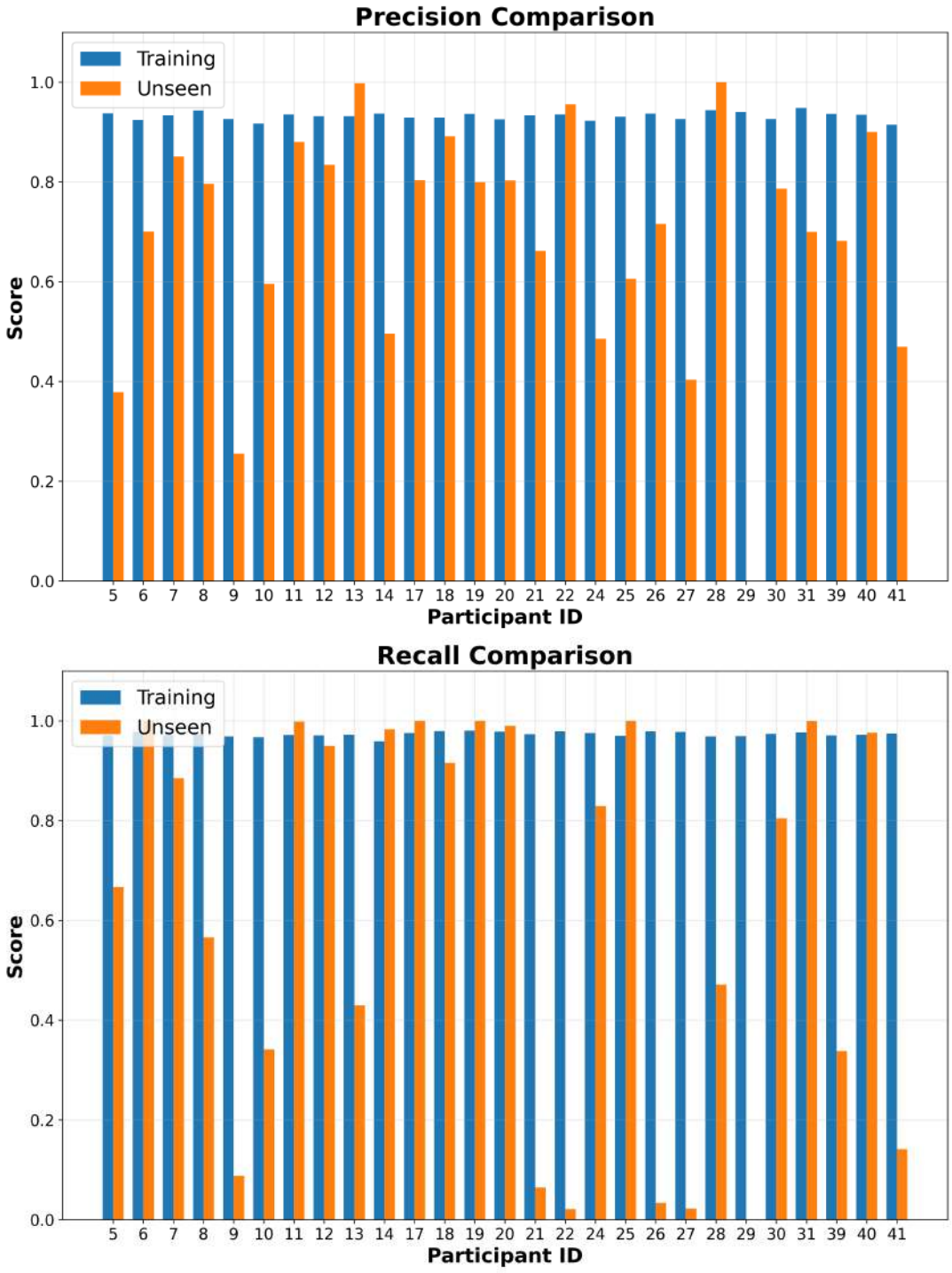


Figure 42. Classification component metrics. *Top:* Precision *Bottom:* Recall

6.3.4 Comprehensive Performance Evaluation

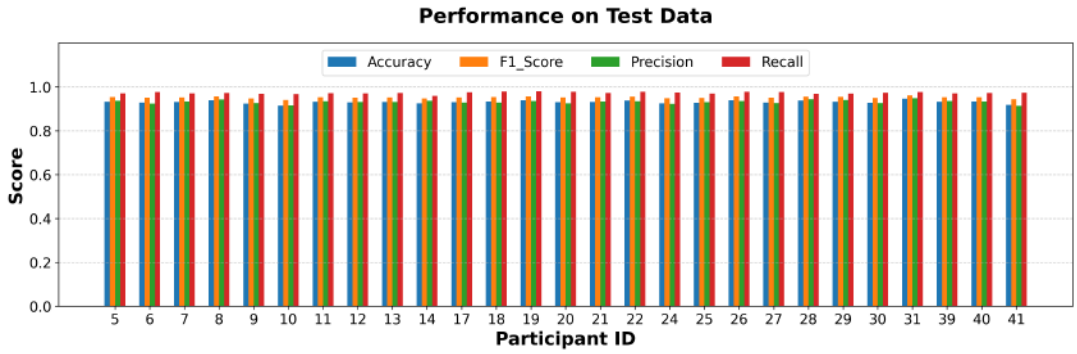


Figure 43. Comprehensive model performance comparison on the test dataset. This unified visualization presents all key classification metrics (accuracy, precision, recall, F1 score, ROC AUC, and PR AUC) across all evaluated models, enabling direct comparison of relative strengths and weaknesses on held-out test data.

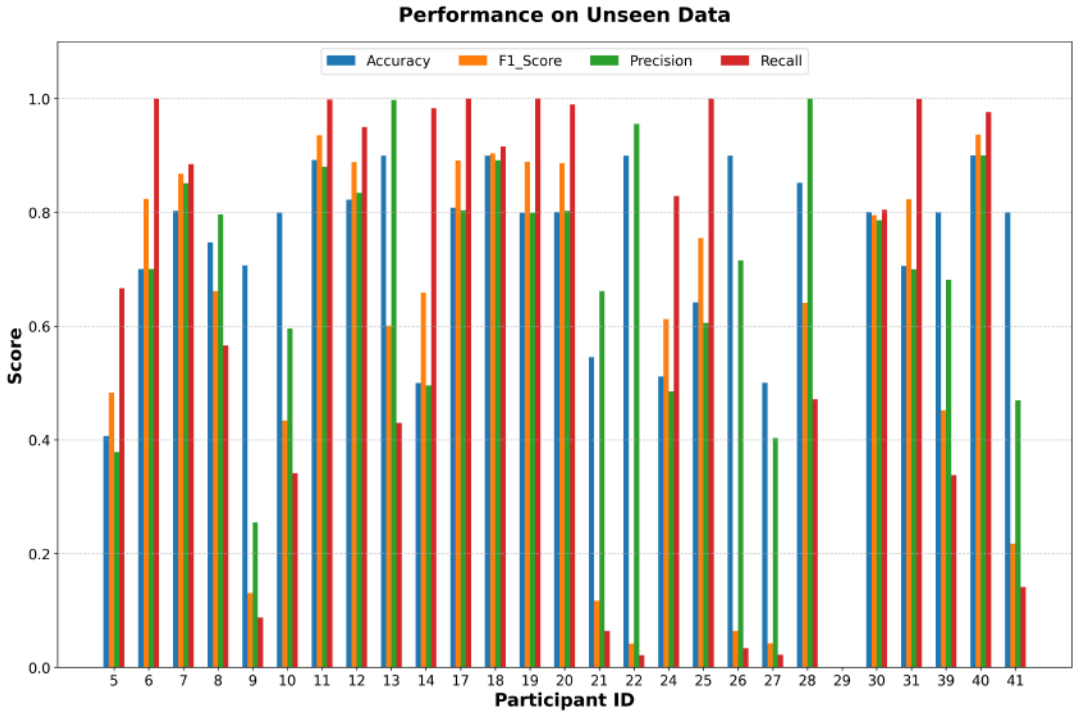


Figure 44. Model robustness evaluation on completely unseen data beyond the training distribution. This critical assessment reveals each model's generalization capabilities in real-world scenarios, highlighting performance stability or deterioration across multiple metrics.

6.3.5 Advanced Performance Insights

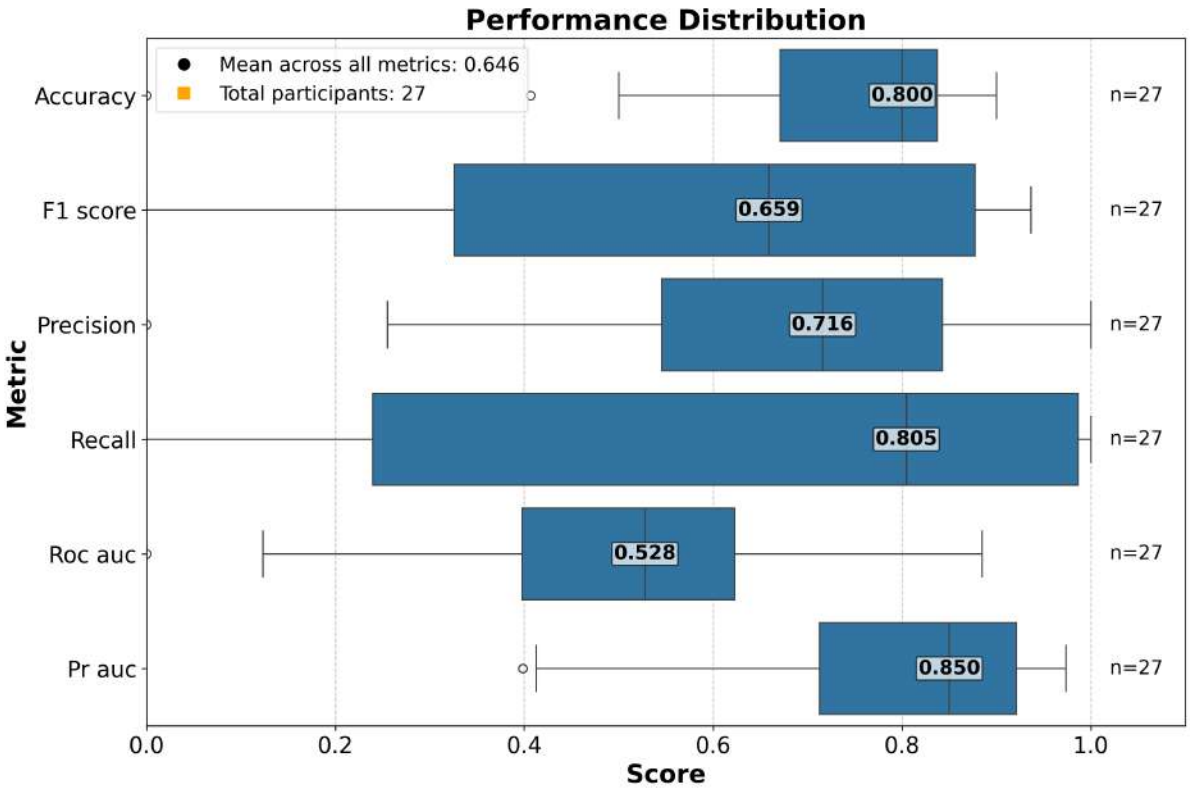
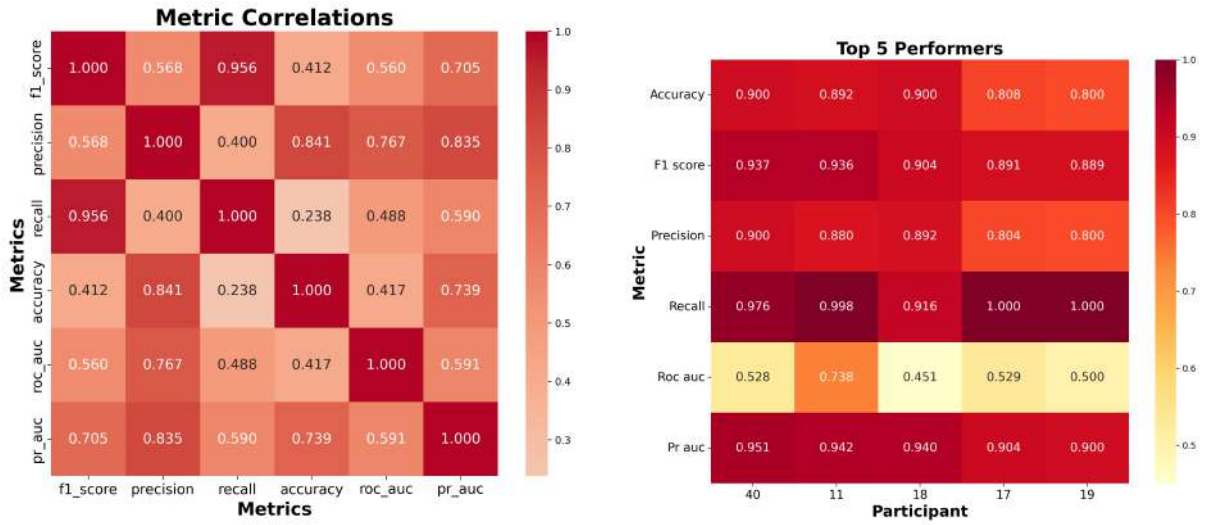


Figure 45. Statistical distribution of performance metrics across all evaluated models. This visualization reveals the central tendency, variance, and potential outliers in model performance, providing insight into the overall effectiveness and consistency of the implemented approach across different evaluation criteria.



(a) Metrics Correlation Analysis **(b)** Top 5 Performing Models

Figure 46. Advanced performance analysis showing (a) correlation patterns between different evaluation metrics, revealing potential redundancies or complementary relationships in the assessment framework; and (b) a focused comparison of the five best-performing models across all metrics, highlighting the leaders in overall classification performance.

6.4 Appendix D: Graphs for Research Question 3 - Participant 18

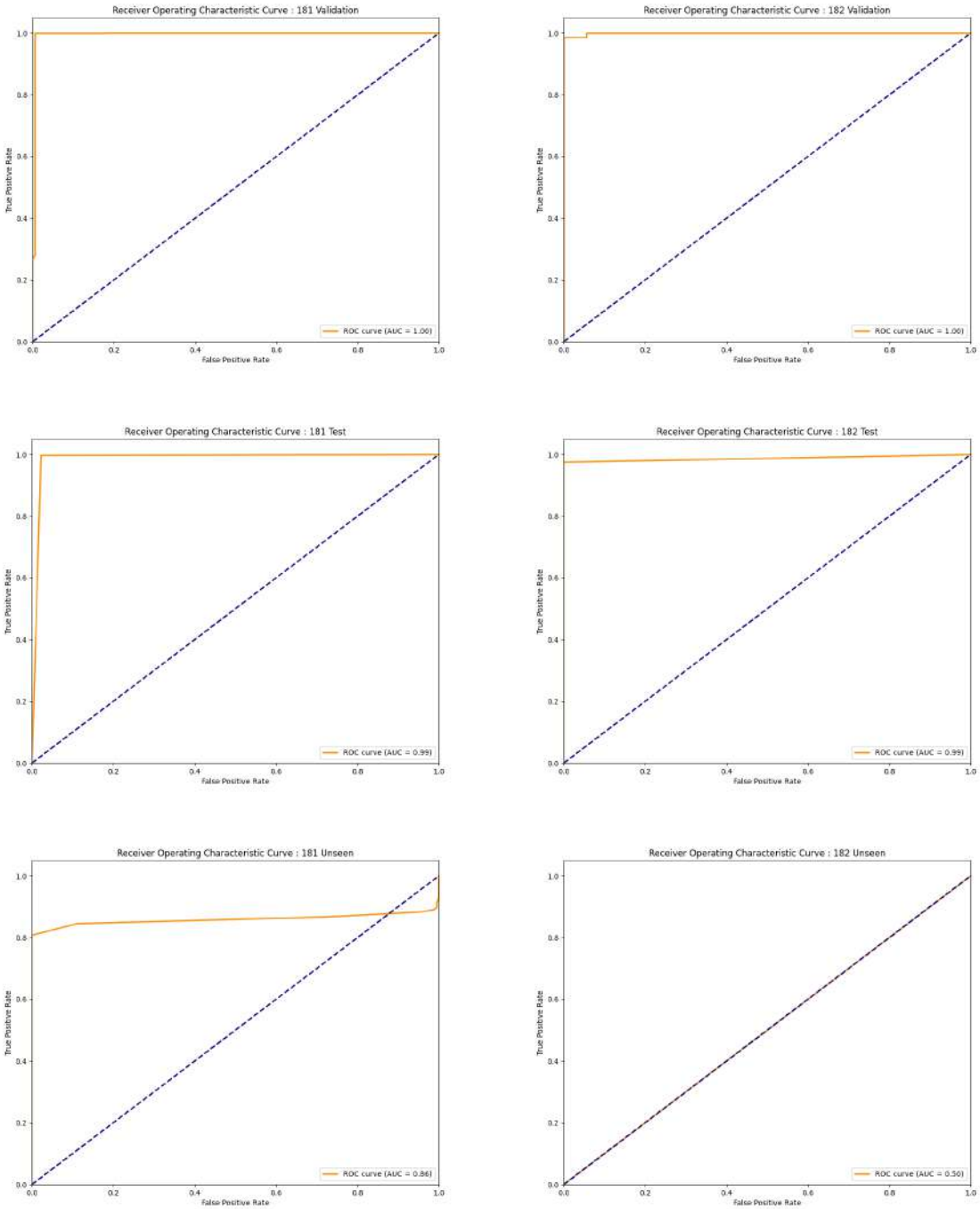


Figure 47. ROC curves for Participant 18 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

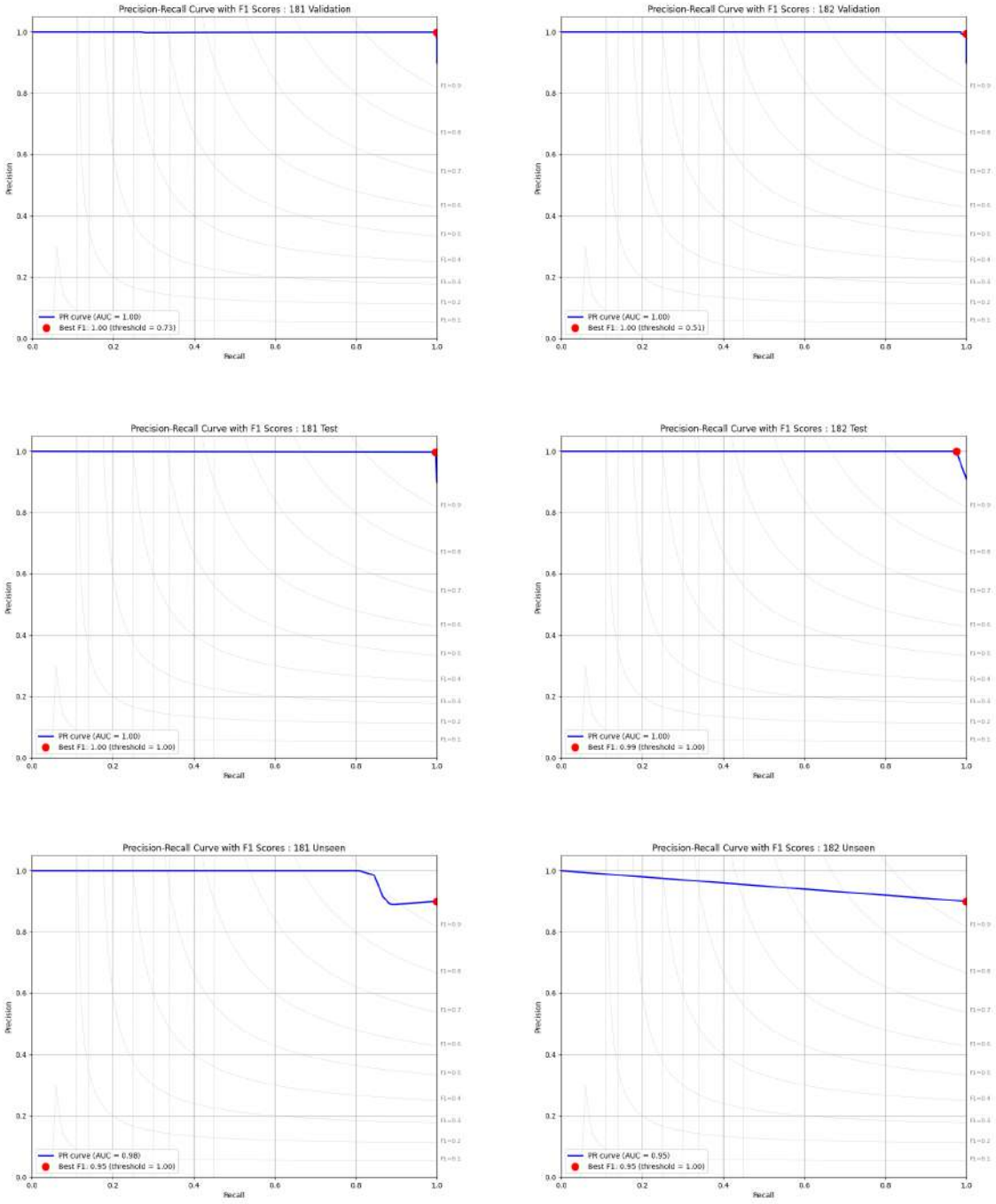


Figure 48. Precision-Recall curves for Participant 18 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

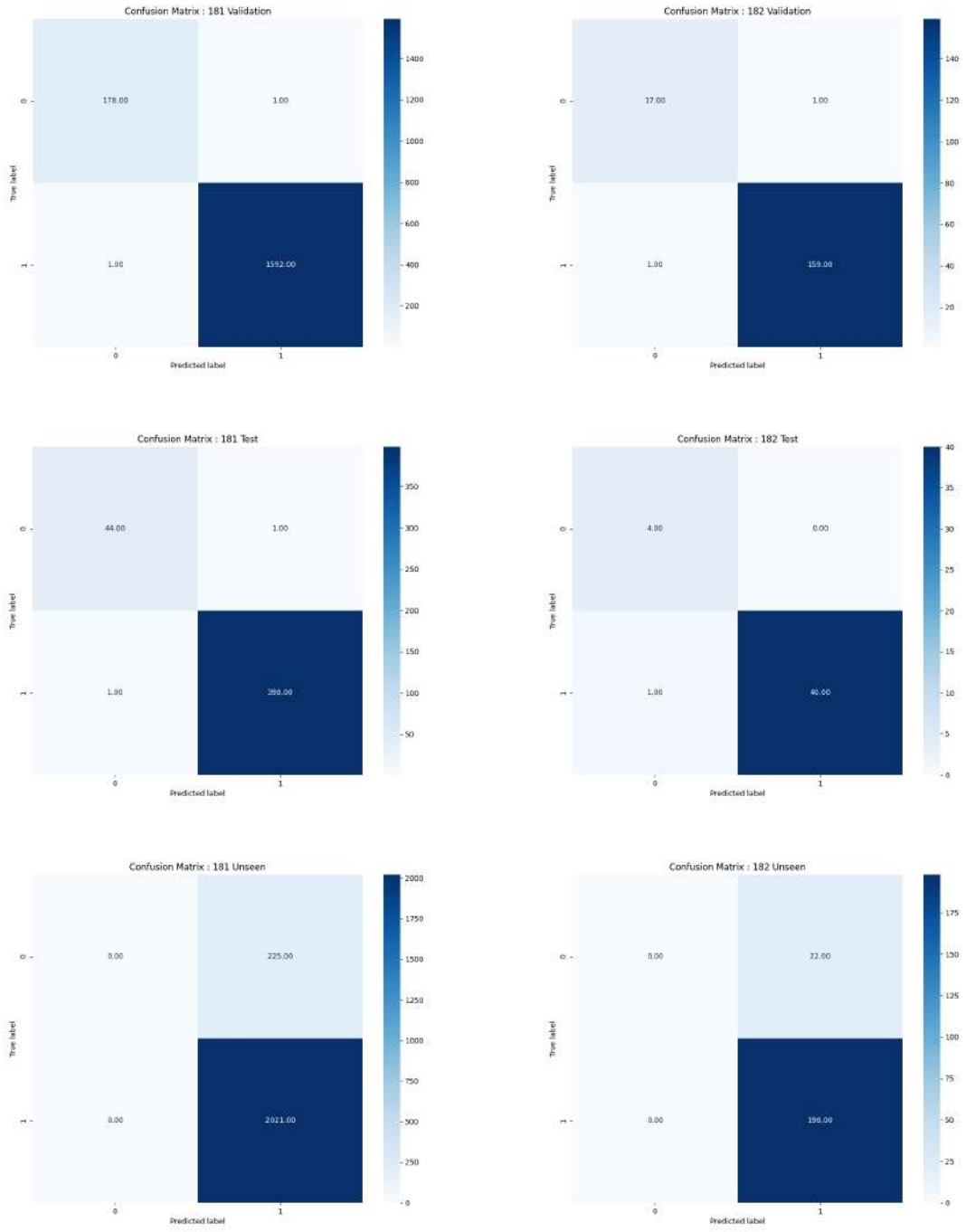


Figure 49. Confusion matrices for Participant 18 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

6.5 Appendix E: Graphs for Research Question 3 - Participant 31

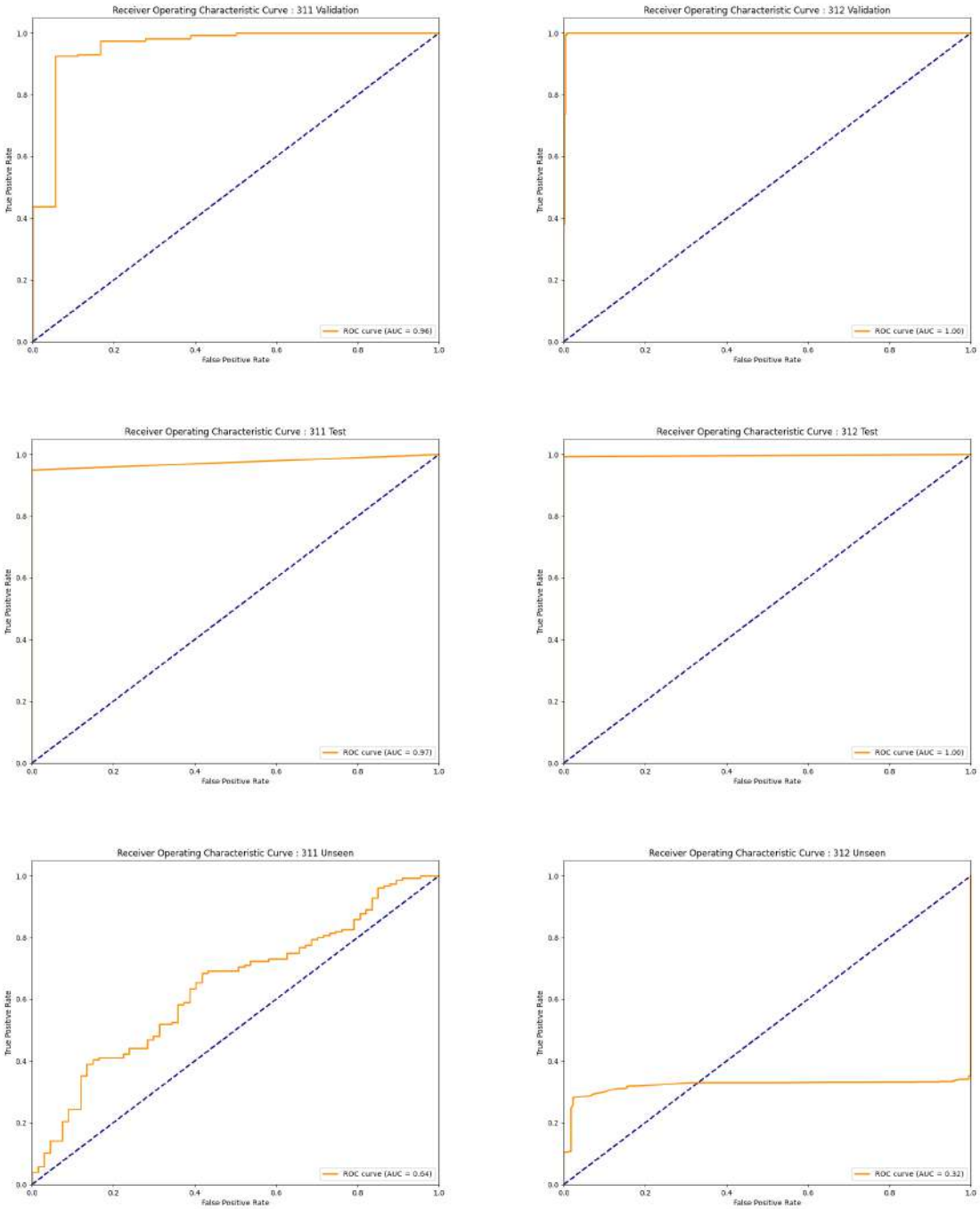


Figure 50. ROC curves for Participant 31 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

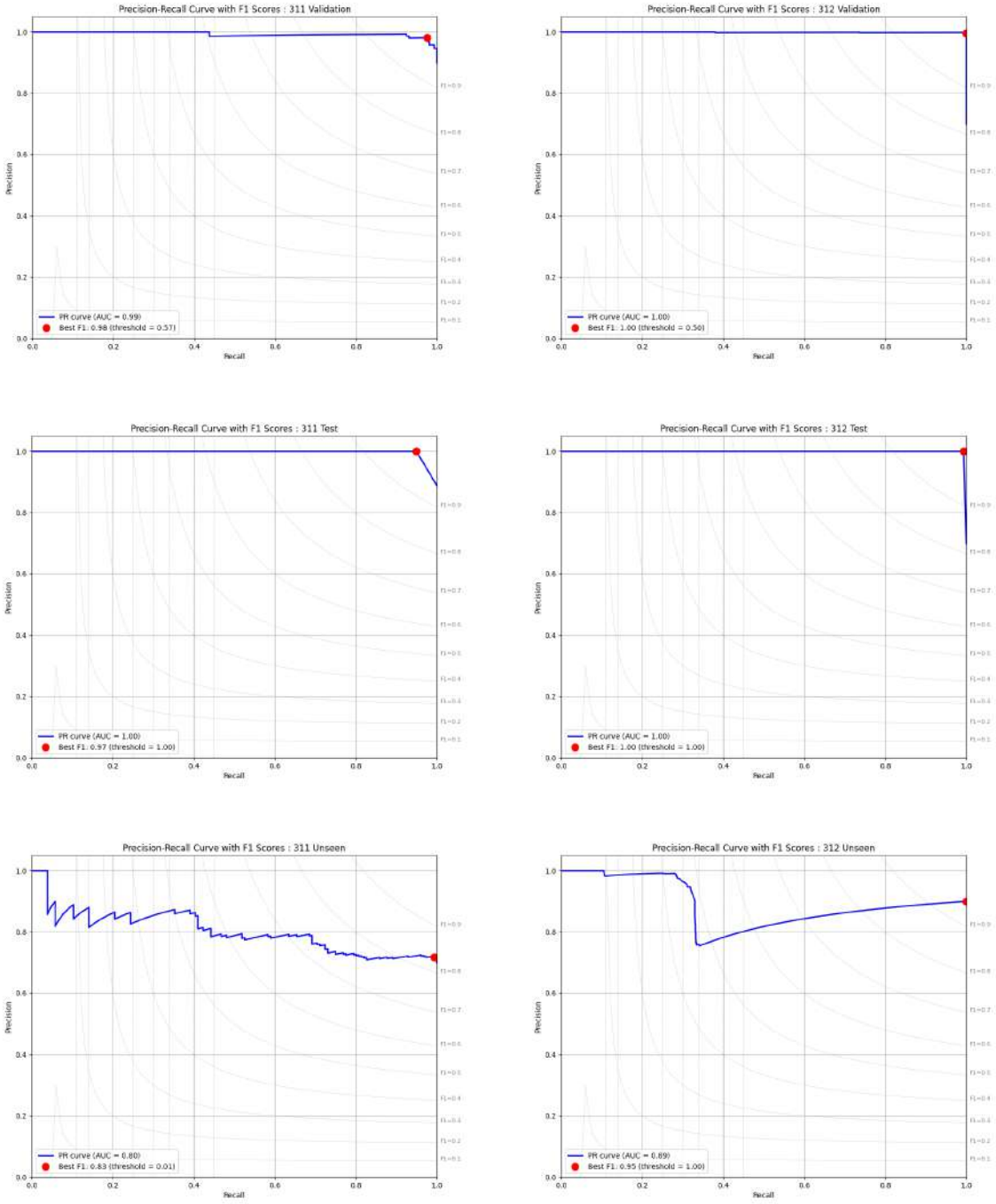


Figure 51. Precision-Recall curves for Participant 31 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

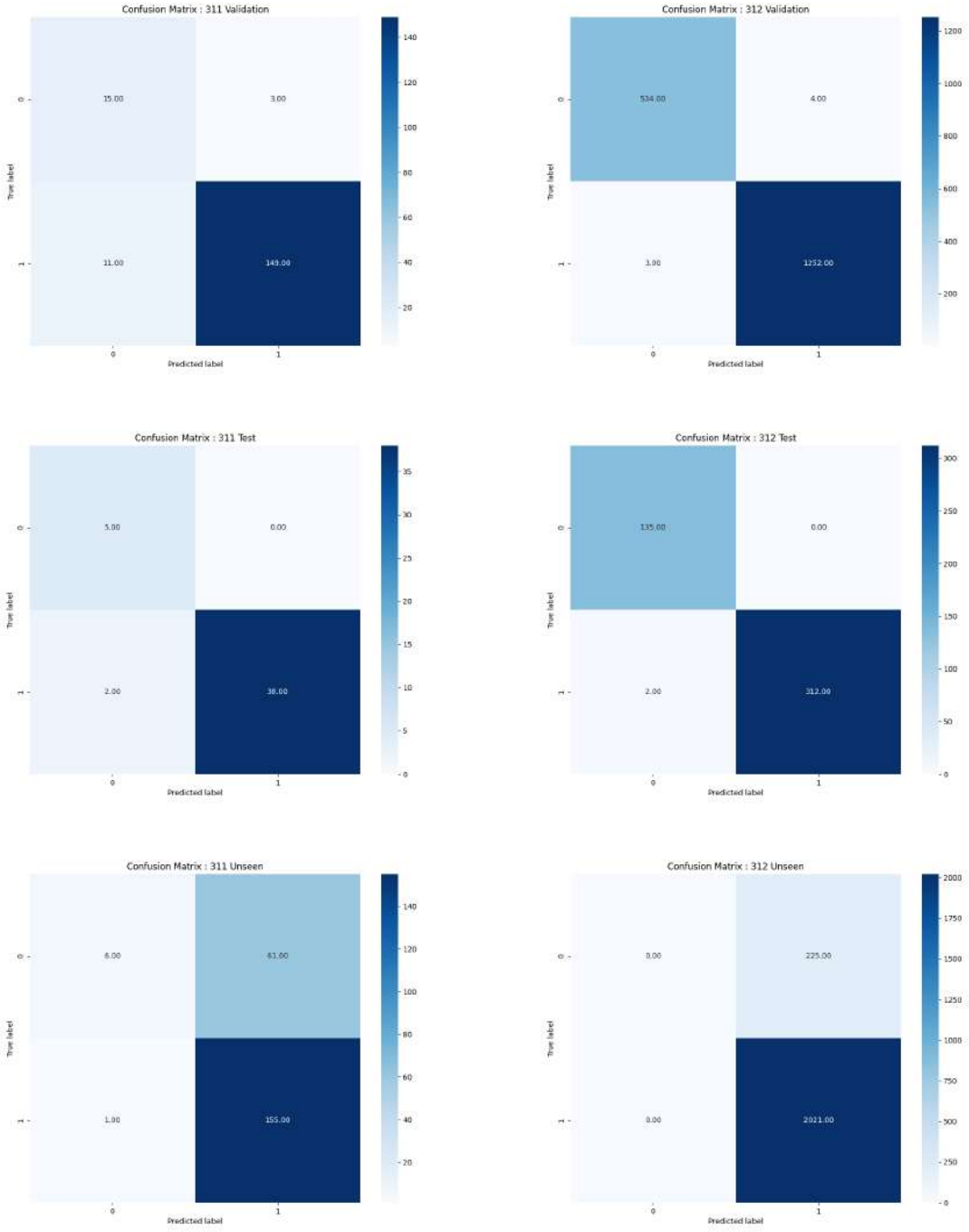


Figure 52. Confusion matrices for Participant 31 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

6.6 Appendix F: Graphs for Research Question 3 - Participant 39

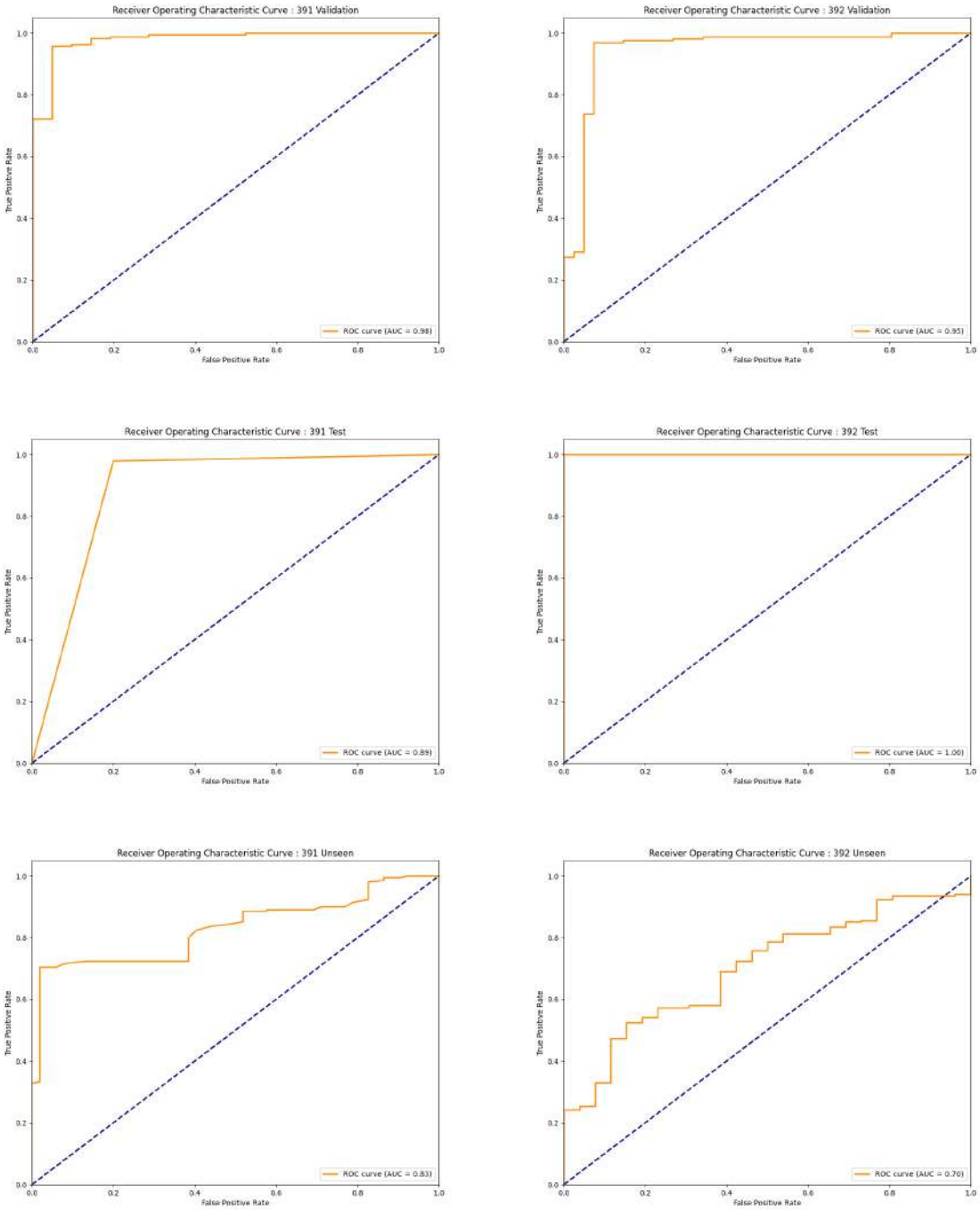


Figure 53. ROC curves for Participant 39 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

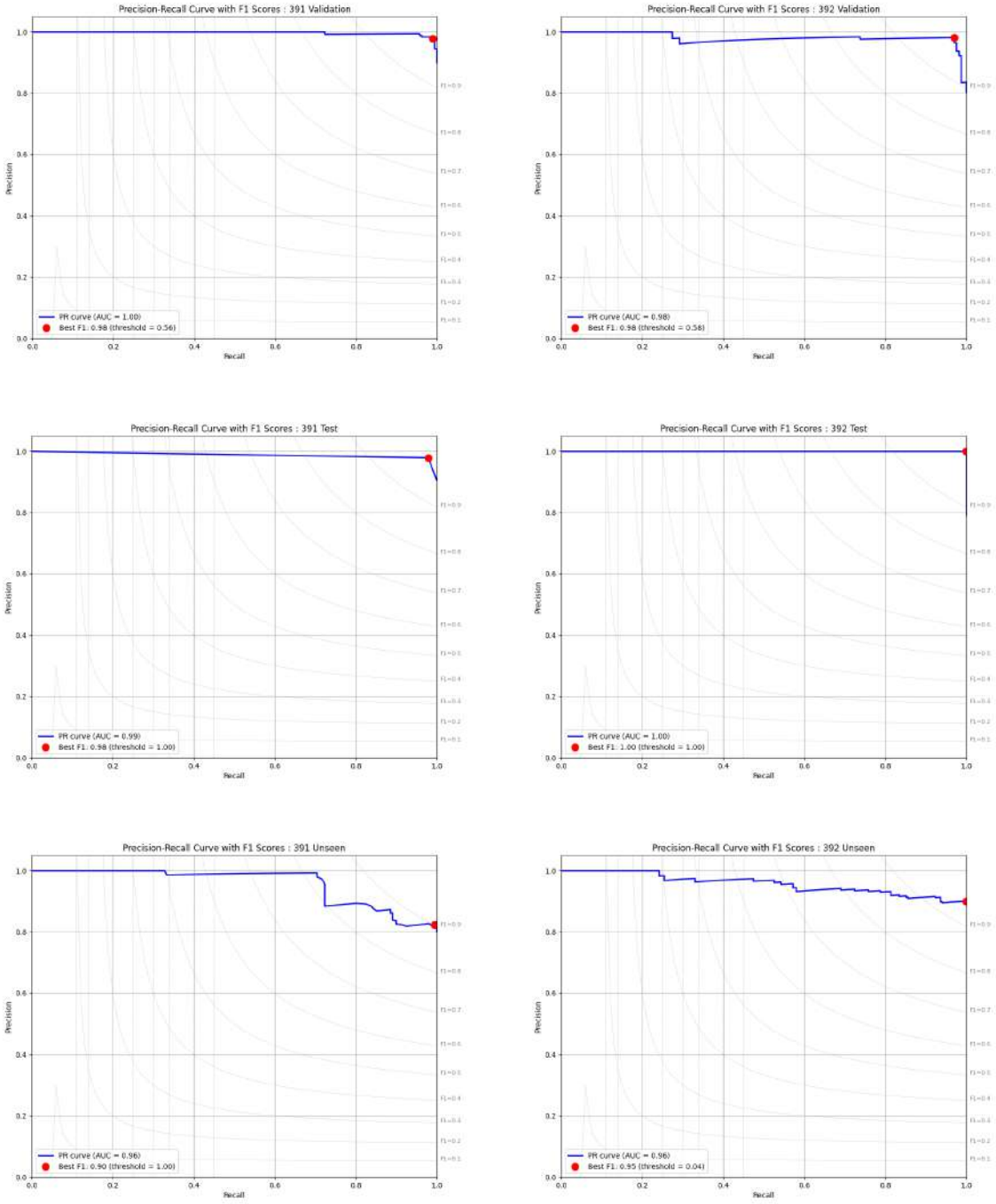


Figure 54. Precision-Recall curves for Participant 39 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

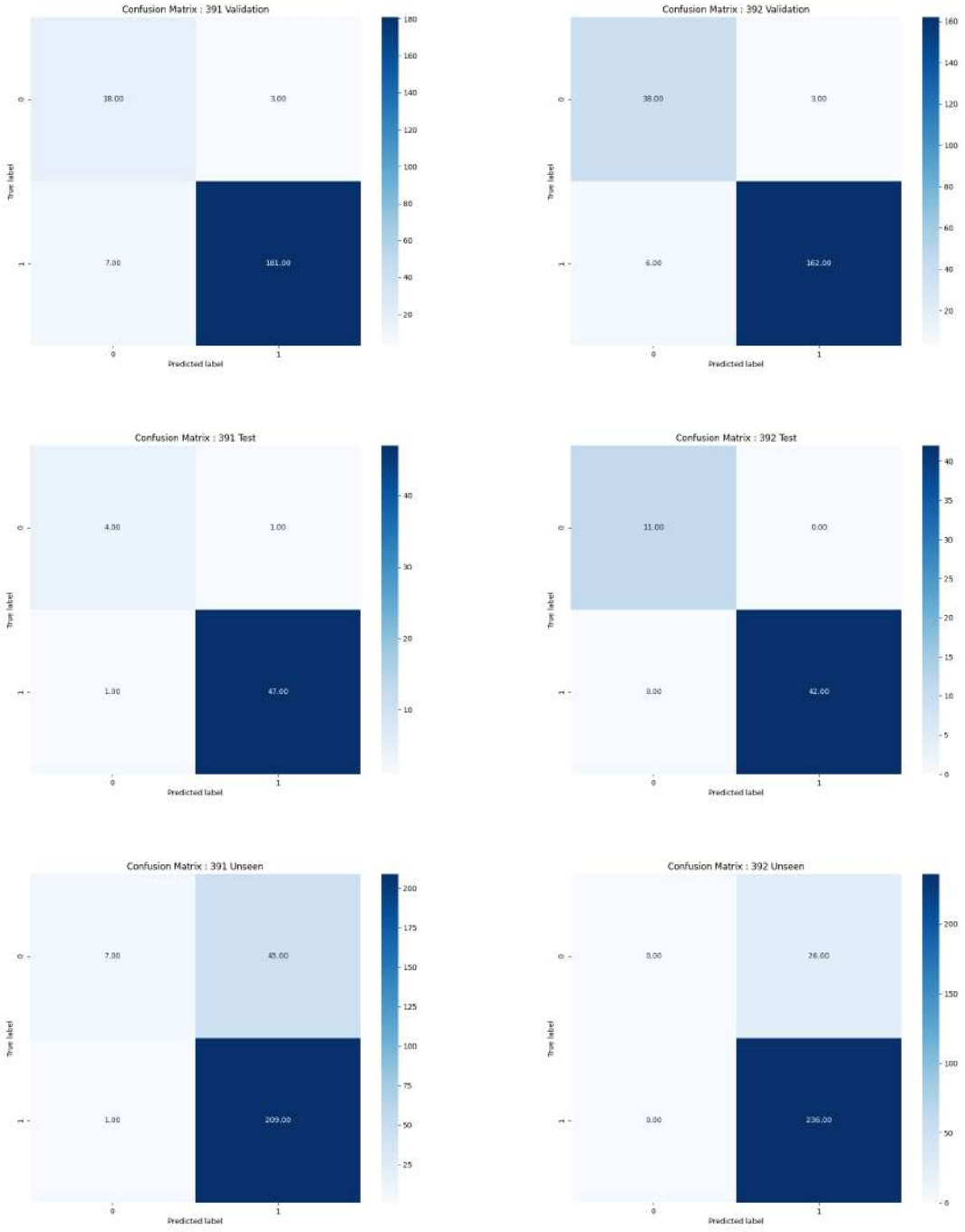


Figure 55. Confusion matrices for Participant 39 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

6.7 Appendix G: Graphs for Research Question 3 - Participant 40

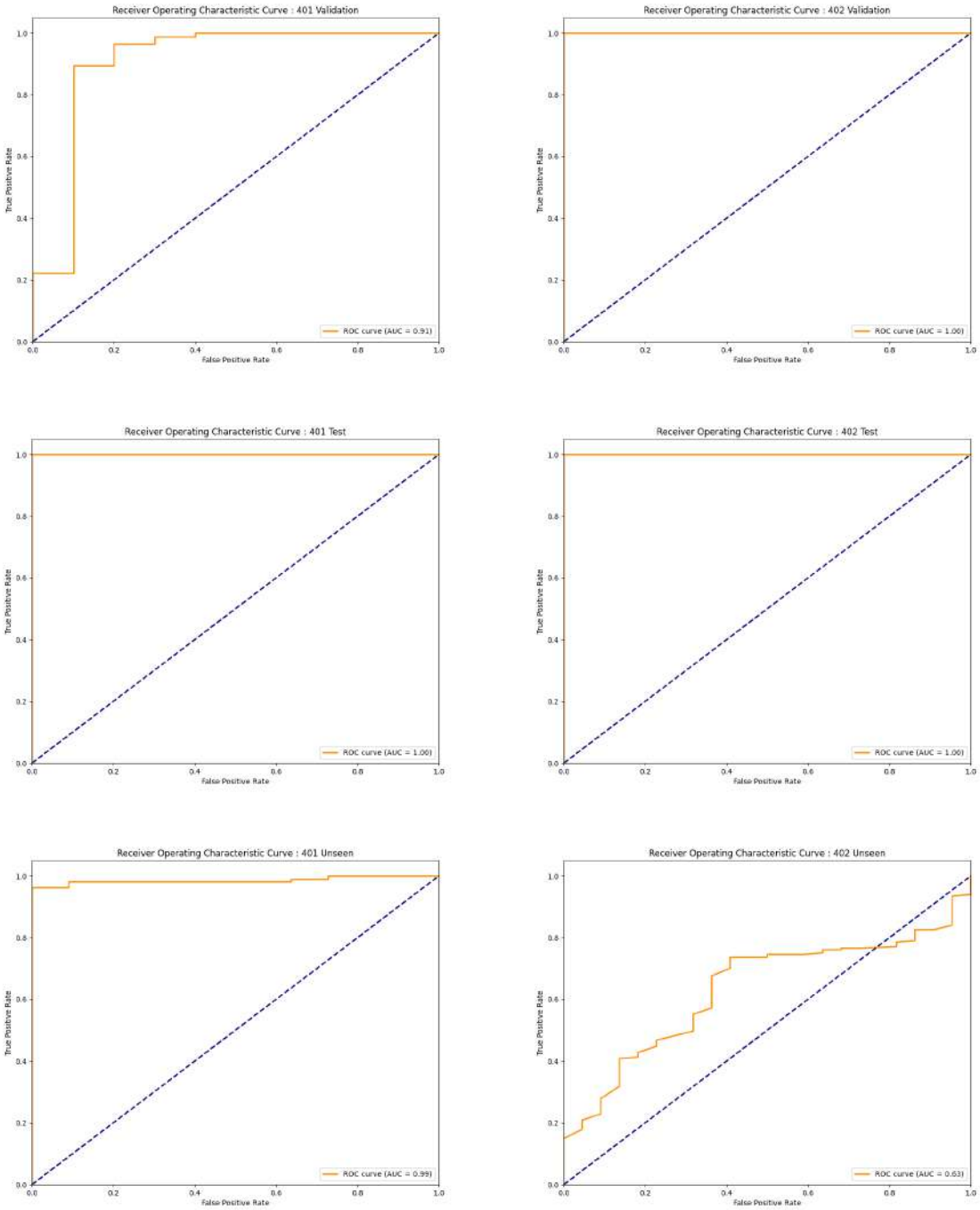


Figure 56. ROC curves for Participant 40 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

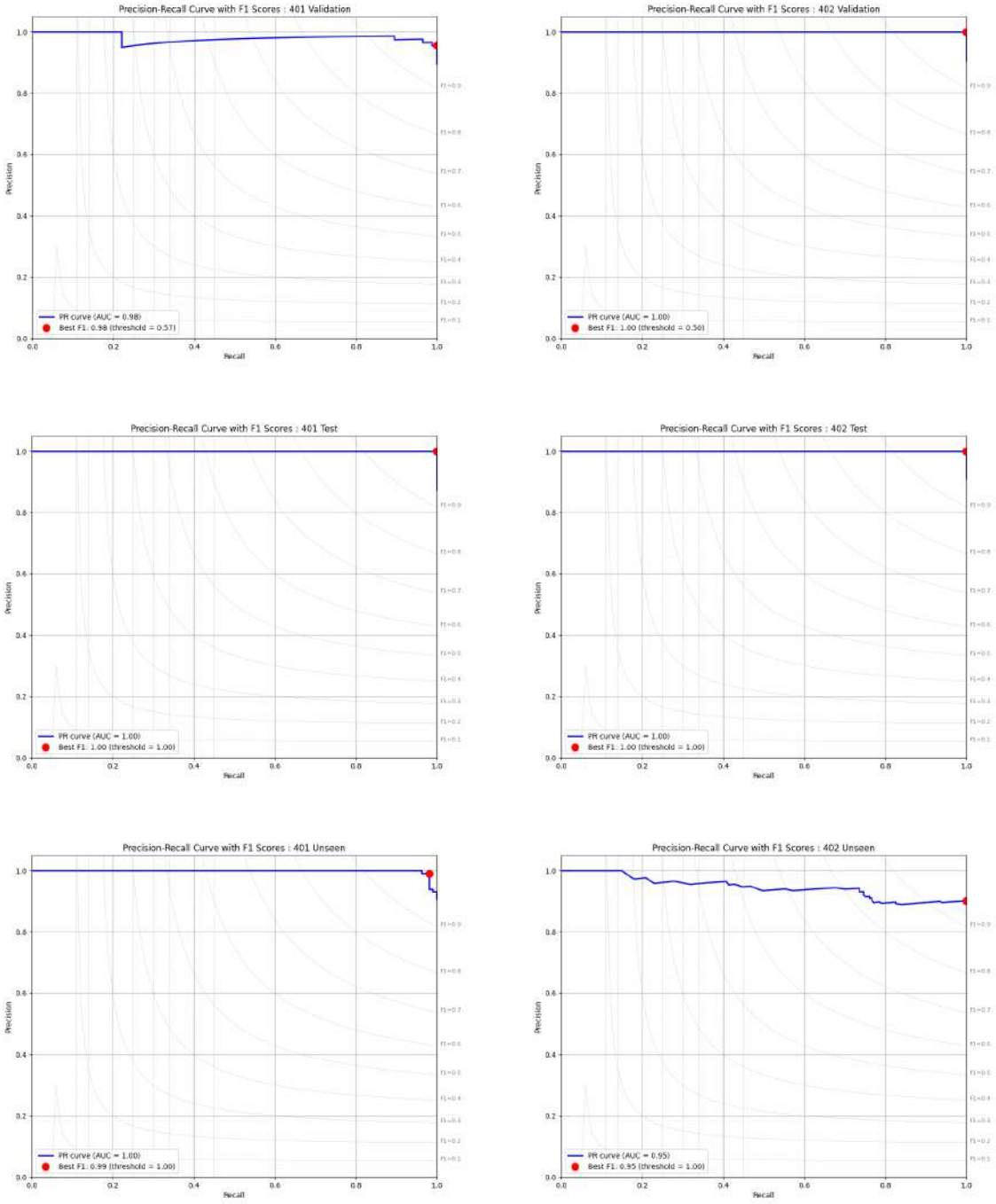


Figure 57. Precision-Recall curves for Participant 40 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

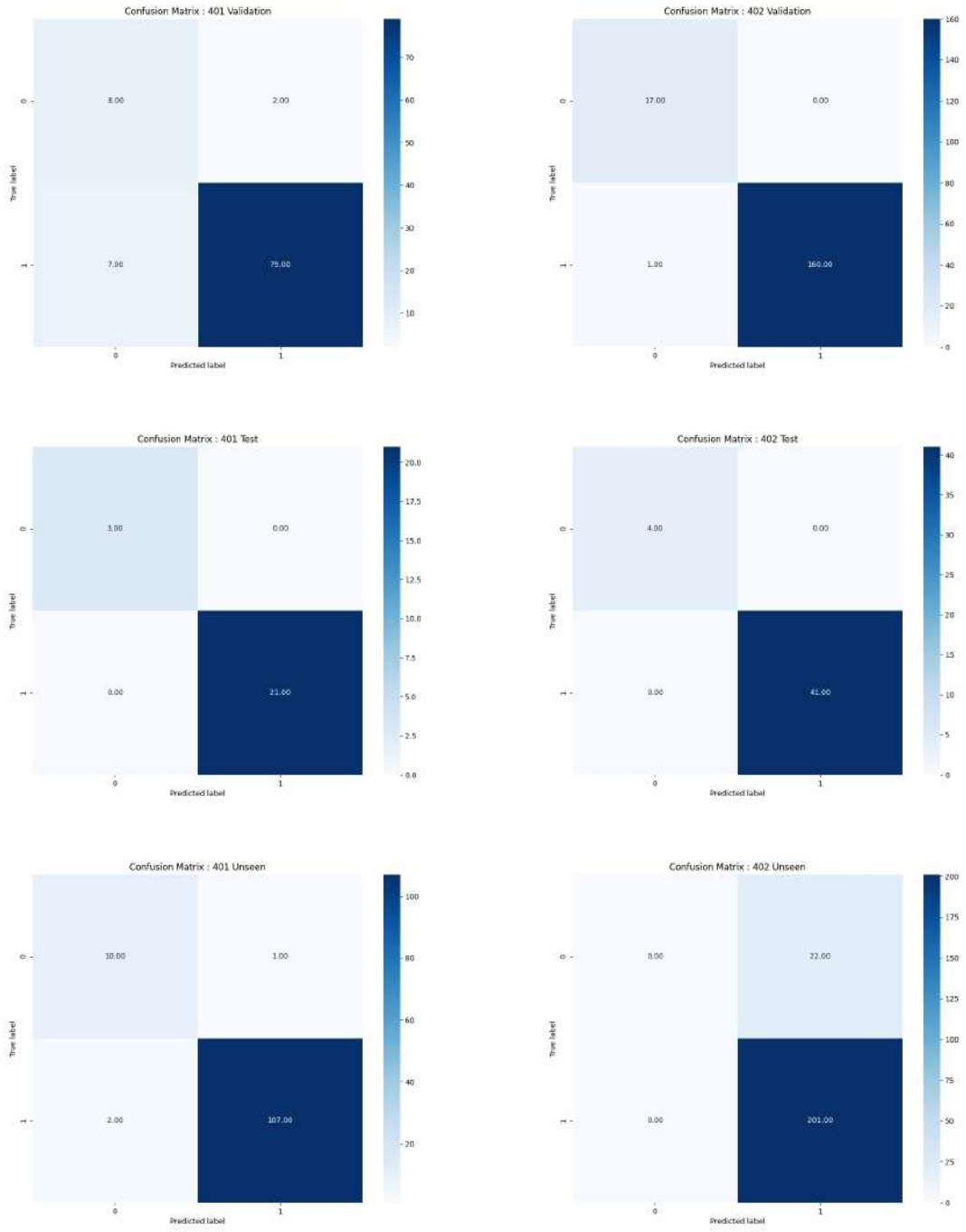


Figure 58. Confusion matrices for Participant 40 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

6.8 Appendix H: Graphs for Research Question 3 - Participant 41

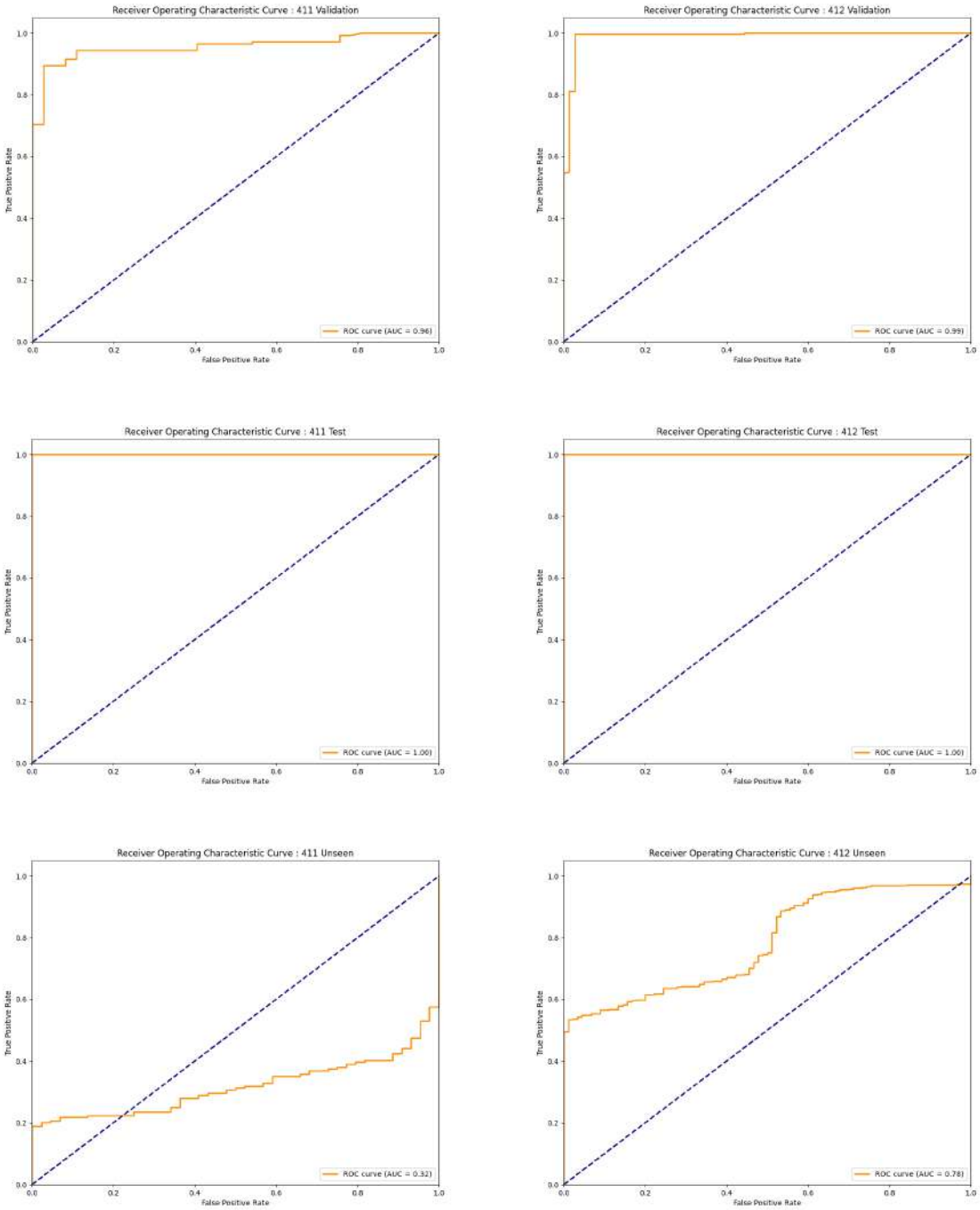


Figure 59. ROC curves for Participant 41 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

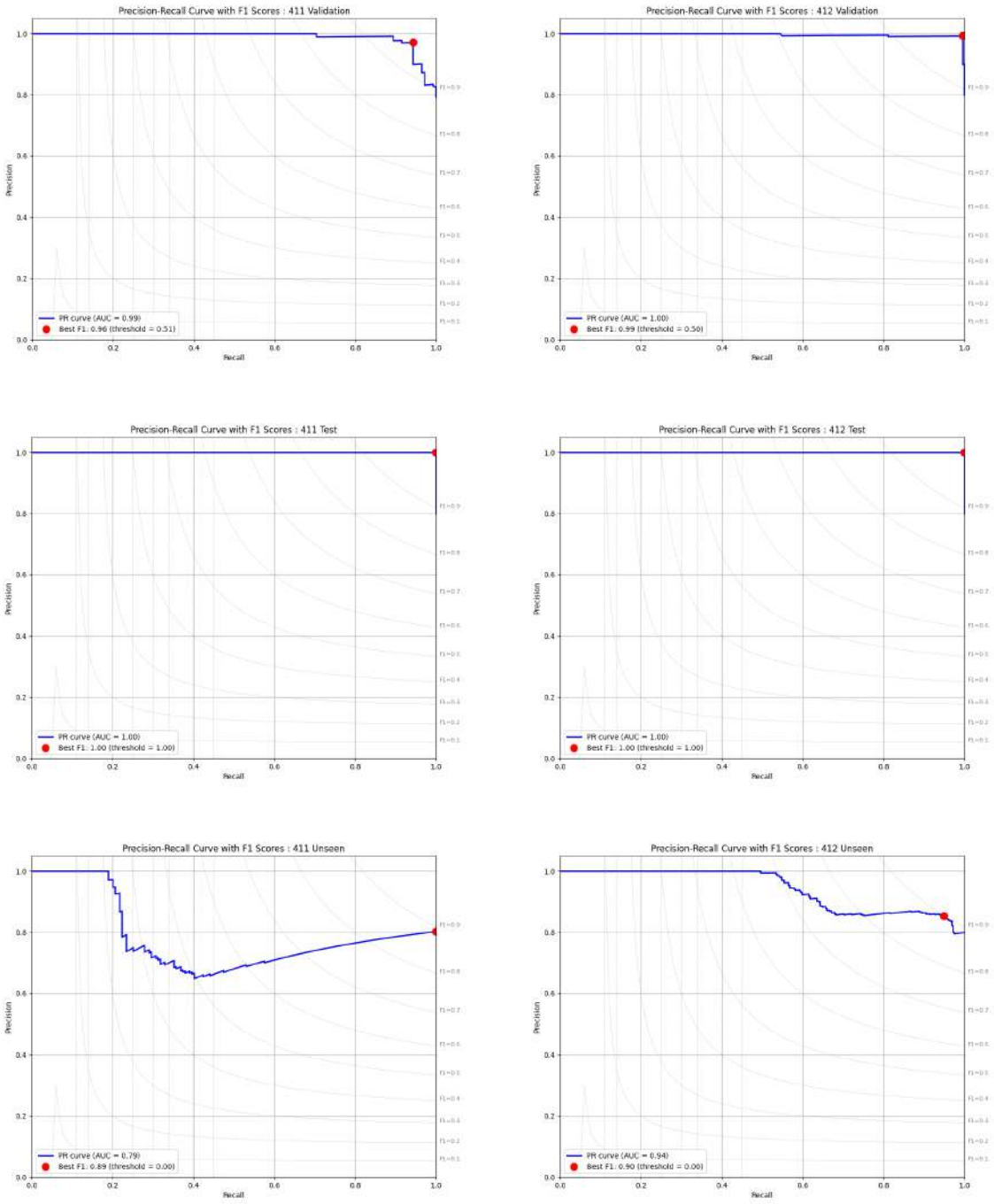


Figure 60. Precision-Recall curves for Participant 41 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

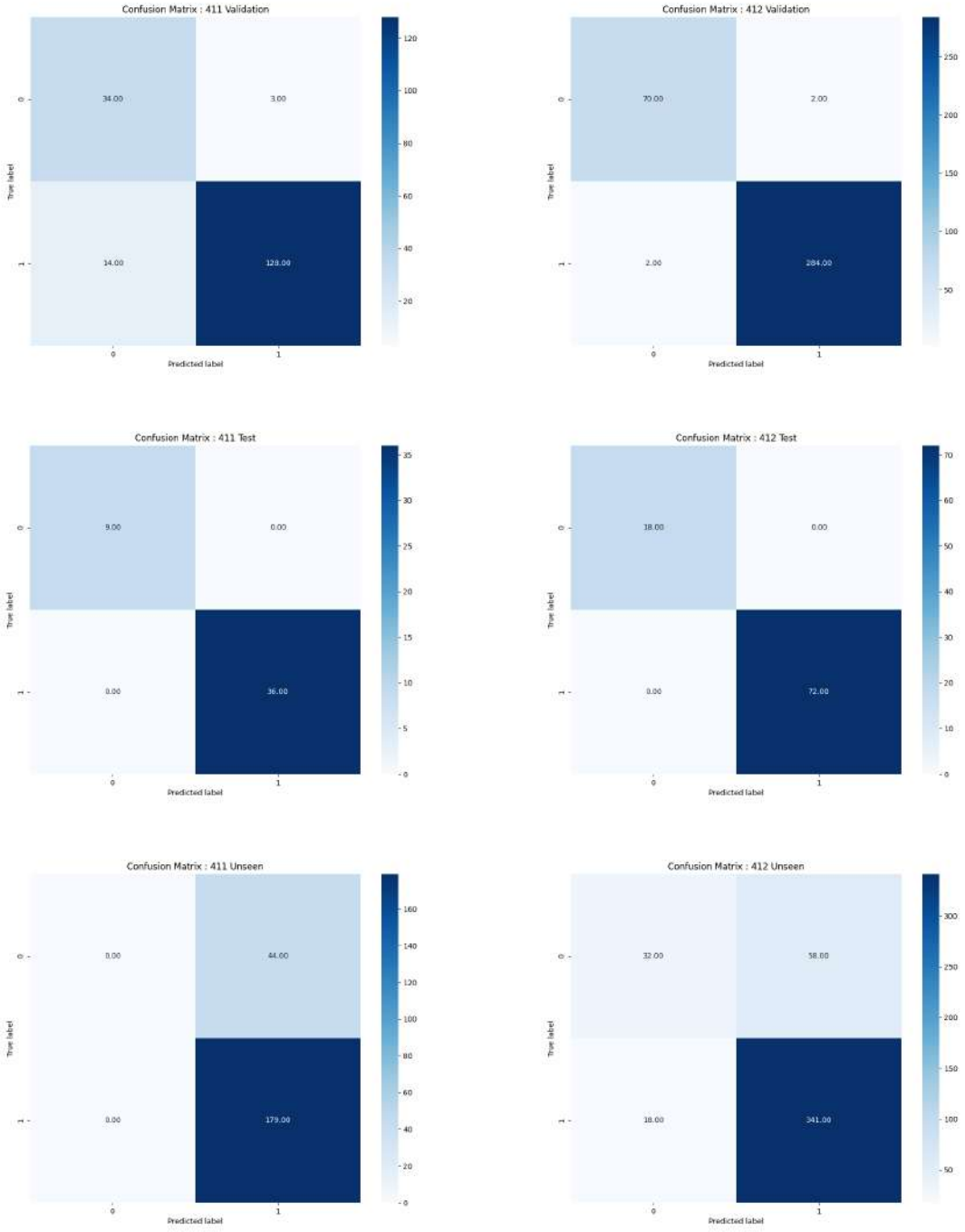


Figure 61. Confusion matrices for Participant 41 across different data sets. *Top row:* Validation data. *Middle row:* Test data. *Bottom row:* Performance on unseen data.

Biographical Sketch

Subigya Gautam, born in Kathmandu, Nepal, pursued a degree in Computer Science at the University of Louisiana at Lafayette, where he developed a deep passion for technology and innovation. His dedication to research led him to join the HCI Lab under Dr. Arun Kulshreshth, focusing on HCI, VR, and ML/DL. Through his work, he explored cutting-edge interactions between humans and technology. His research and commitment culminated in earning a master's degree in Computer Science from the University of Louisiana at Lafayette in Spring 2025. His academic journey reflects a strong drive for innovation, contributing to advancements in human-computer interaction.