

Speech Features per i task di Speaker Identification e Verification

Autore: Gabriele Nicolò Costa

Outline

1

Introduzione al
problema

2

Speech features
e speech
embeddings

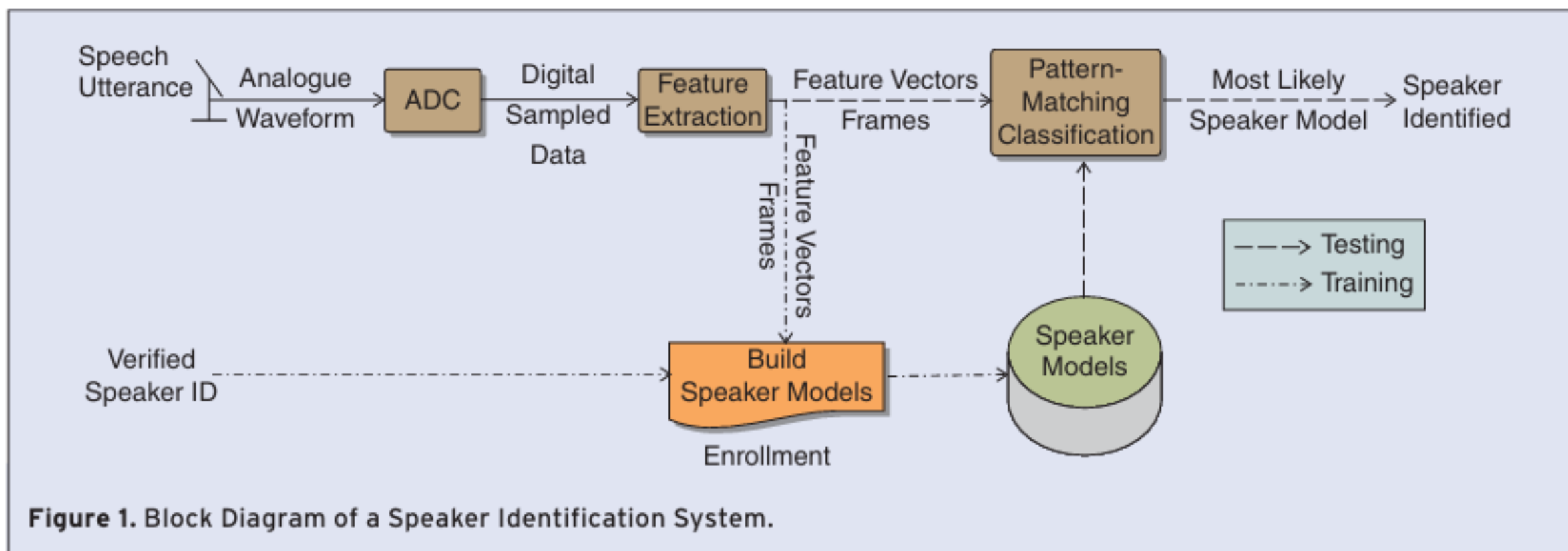
3

Speaker Models
e Backend –
Architettura
proposta

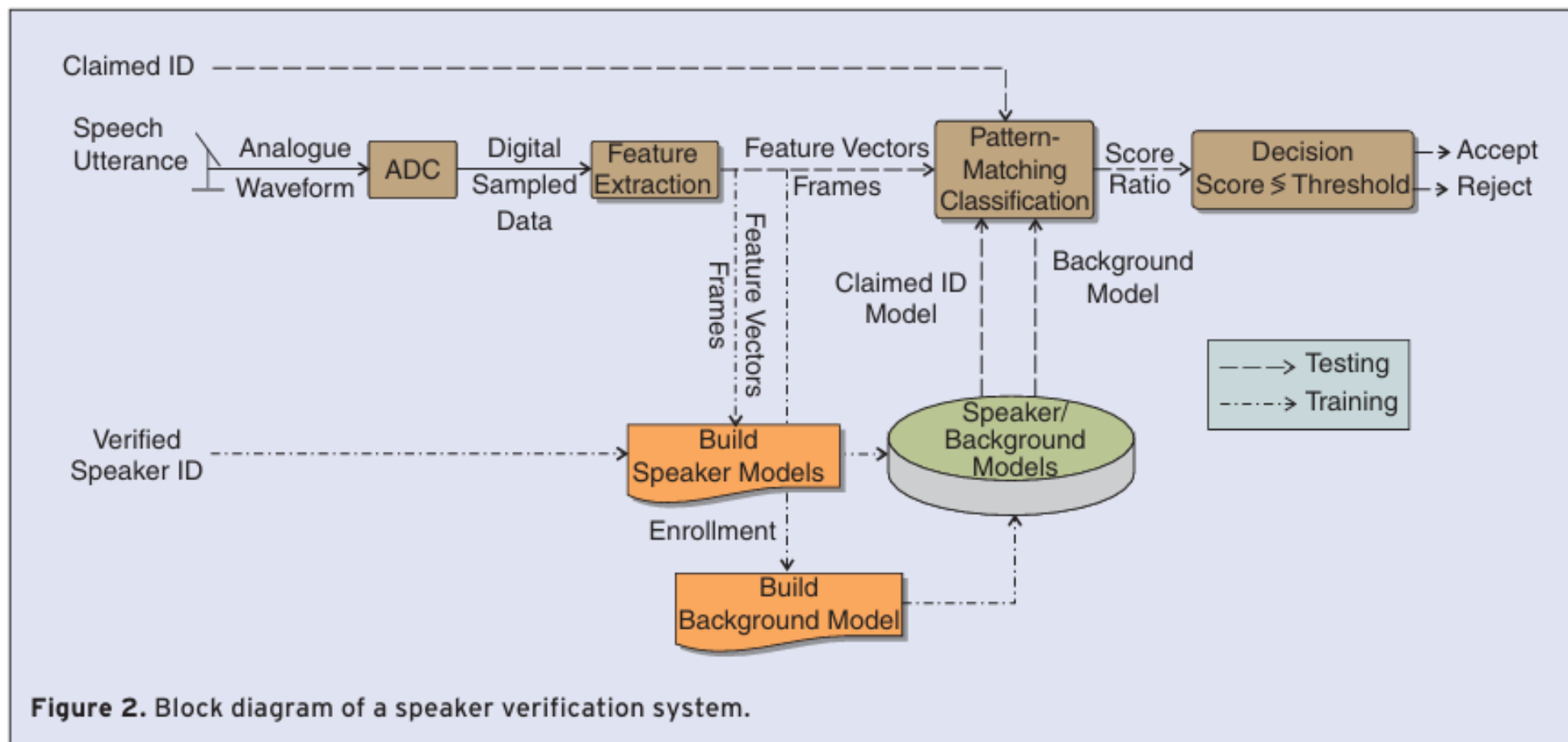
4

Risultati
sperimentali

Speaker Identification



Speaker Verification



Metriche e Valutazione dei sistemi

Task	Obiettivo	Metriche utilizzate
Speaker Identification (Closed-set)	Riconoscere l'identità dello speaker tra un insieme noto di speakers (classificazione multiclasse).	<ul style="list-style-type: none">• Accuracy• Precision, Recall, F1-score (per classe e macro media)
Speaker Verification (Open-set)	Verificare se uno speaker è chi dichiara di essere (classificazione binaria: accetta/rifiuta).	<ul style="list-style-type: none">• True Positive Rate (TPR)• False Positive Rate (FPR)• False Acceptance Rate (FAR)• False Rejection Rate (FRR)• Equal Error Rate (EER)

Table 2.1: Task, obiettivi e metriche nei sistemi di speaker recognition

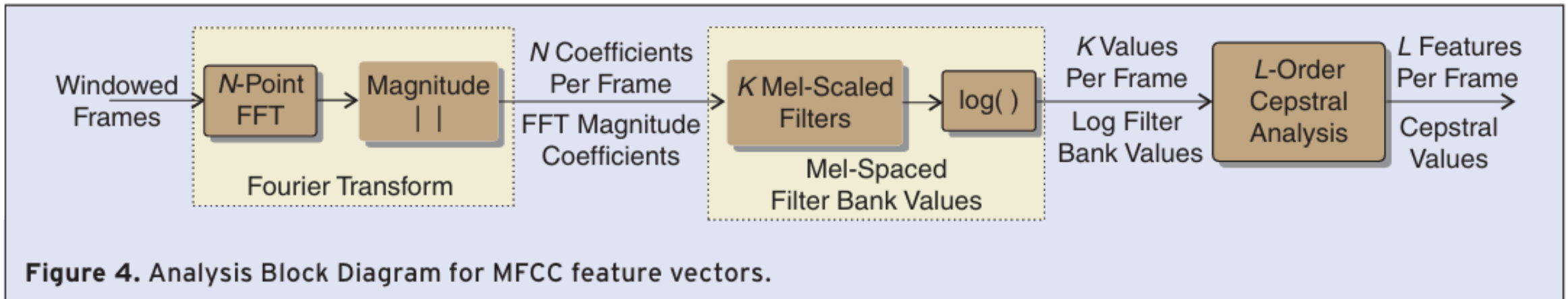
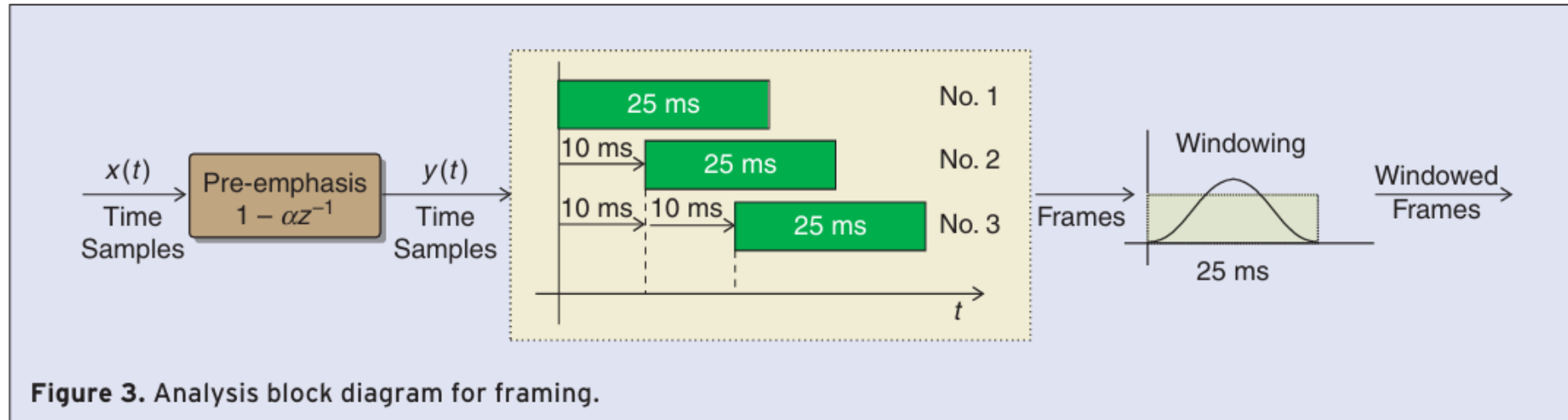
Speech Features

Cepstrum

FilterBanks e Mel-scale

Mel-Frequency Cepstral
Coefficients (MFCCs)

Framework estrazione Features



Speech Embeddings

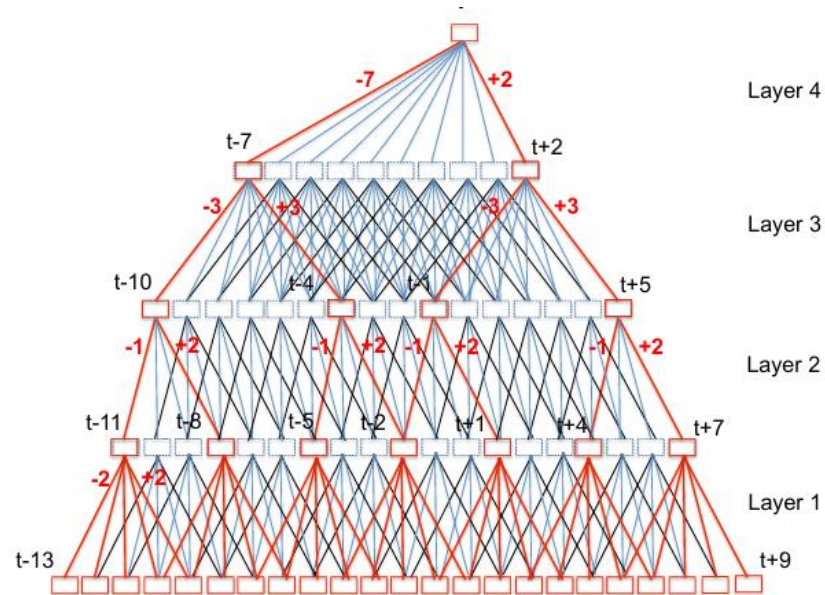


Figure 1: Computation in TDNN with sub-sampling (red) and without sub-sampling (blue+red)

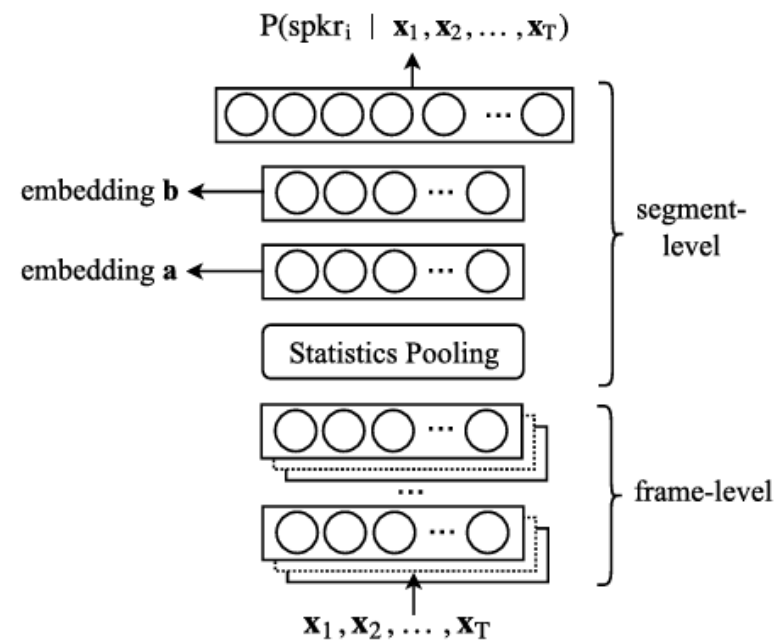


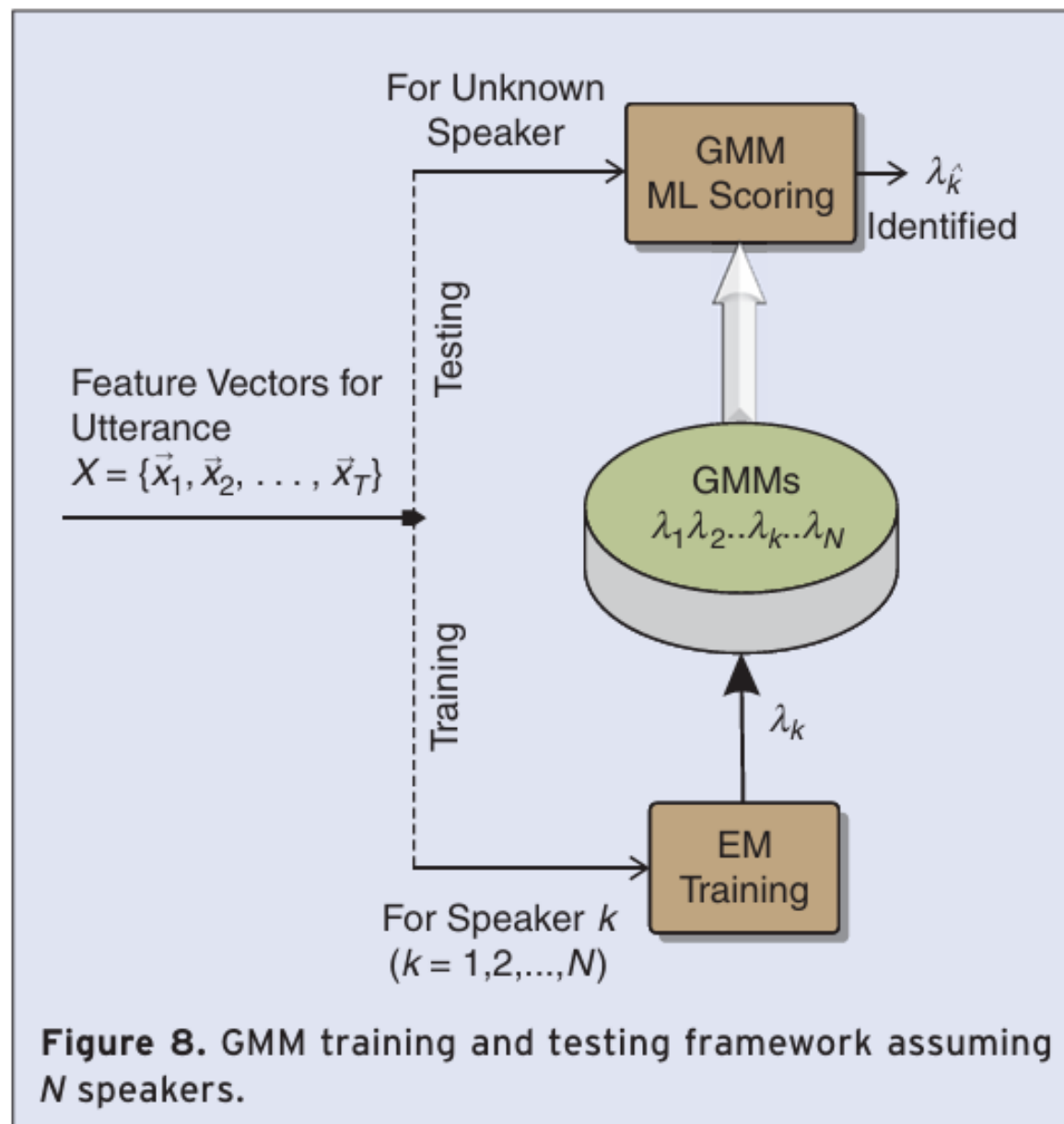
Figure 1: Diagram of the DNN. Segment-level embeddings (e.g., **a** or **b**) can be extracted from any layer of the network after the statistics pooling layer.

Speaker Identification and Verification pipeline

TABLE I: COMPARATIVE STUDY OF VARIOUS SPEAKER VERIFICATION SYSTEM BASED ON DNN ARCHITECTURES

Reference	Type of System	Input Features	DNN Type	Score Function	Baseline System	Dataset	Score (%EER)
[10]	Text-dependent	40 log Mel-filter bank coefficients	7-layered, fully-connected	PLDA	UBM/i-vector	NIST SRE'12, Noisy narrowband	1.39
[12]	Text-independent	39 dimension PLP	7-layered RBM	Cosine Distance	GMM-UBM	NIST SRE'05-06	0.88
[14]	Text-independent	60 dimension MFCCs	5-layered	PLDA	UBM/i-vector	NIST SRE'12, Switchboard I, II, III	1.58
[16]	Text-dependent	39 dimension PLP	4-layered, fully-connected	Cosine Distance	UBM/i-vector	Self-created	1.21
[17]	Text-dependent	20 dimension MFCCs	4-layered, fully-connected	PLDA	UBM/i-vector, GMM-DTW	RSR 2015 Specifically designed for text dependent speaker verification.	0.2
[18]	Text-dependent	39 dimension PLP	4-layered, fully-connected	PLDA	GMM-UBM, d-vector, j-vector	SR 2015	0.54
[19]	Text-dependent	40 dimension MFCCs	7-layered, multi-splice time delay	GPLDA	GMM-UBM	NIST SRE'10	7.2
[20]	Text-independent	Phone-blind & Phone-aware 40 dimensional d-vectors	7-layered, time-delay	PLDA	UBM/i-vector	Fisher dataset, CSLT-CUDGT2014	8.37
[21]	Text-independent	20 dimension MFCCs	4-layered, temporal pooling	PLDA	UBM/i-vector	US English telephonic speech,	5.3

Speaker Models e Background models



Architettura proposta

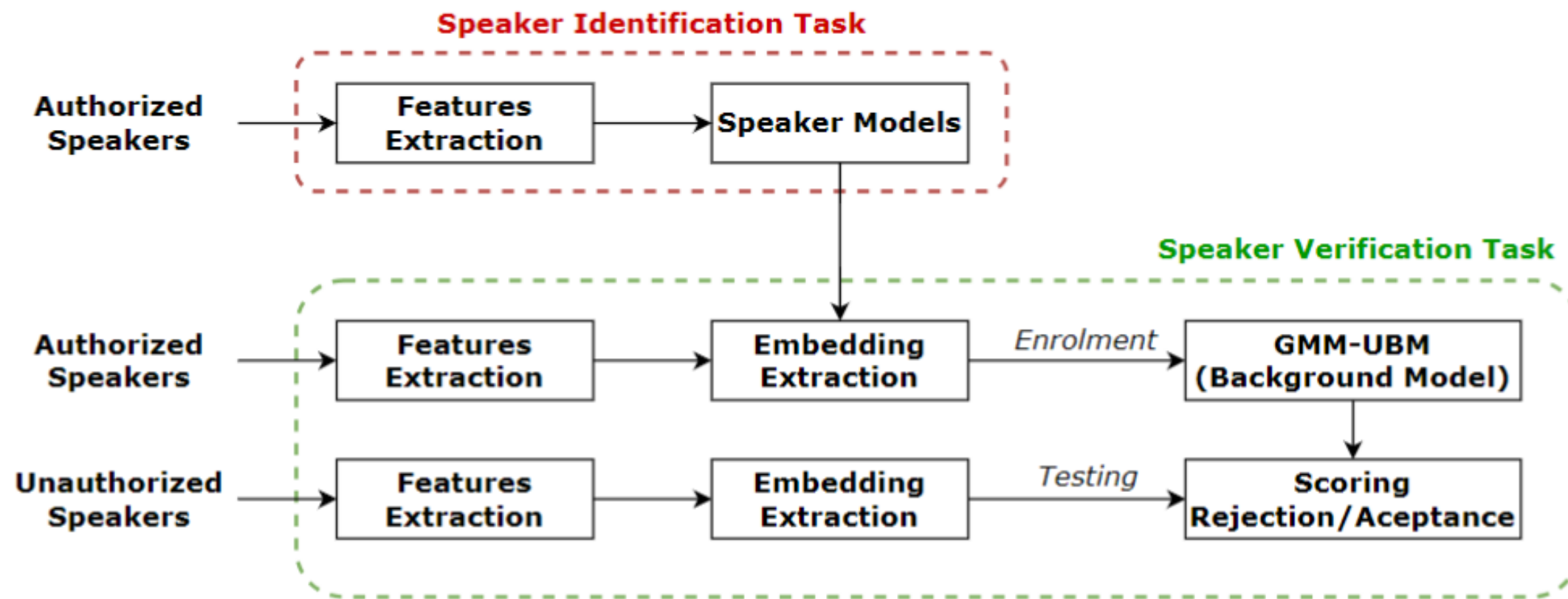


Figure 4.2: Architettura del progetto proposta

Risultati sperimentali

<i>Feature Extraction</i>		<i>Speaker Identification</i>	<i>Speaker Verification</i>	
Features used	Frame Analysis	Model Used	Embedding	Backend
25 MFCCs	25ms frame length 10ms frame hop	DNN	Bottle-neck layer	GMM-UBM
25 MFCCs	25ms frame length 10ms frame hop	DNN	Last Hidden layer	GMM-UBM
25 MFCCs	25ms frame length 10ms frame hop	RNN	Mean Hidden Layers	GMM-UBM
25 MFCCs	25ms frame length 10ms frame hop	CNN+LSTM	Mean Hidden Layers	GMM-UBM
20 MFCCs	25ms frame length 10ms frame hop	I-DNN	Segment Embedding	GMM-UBM
24 MFCCs, 24 Delta-MFCCs VAD 30% energy	25ms frame length 10ms frame hop	TDNN	X-embedding	GMM-UBM

Table 5.1: Riassunto degli esperimenti condotti

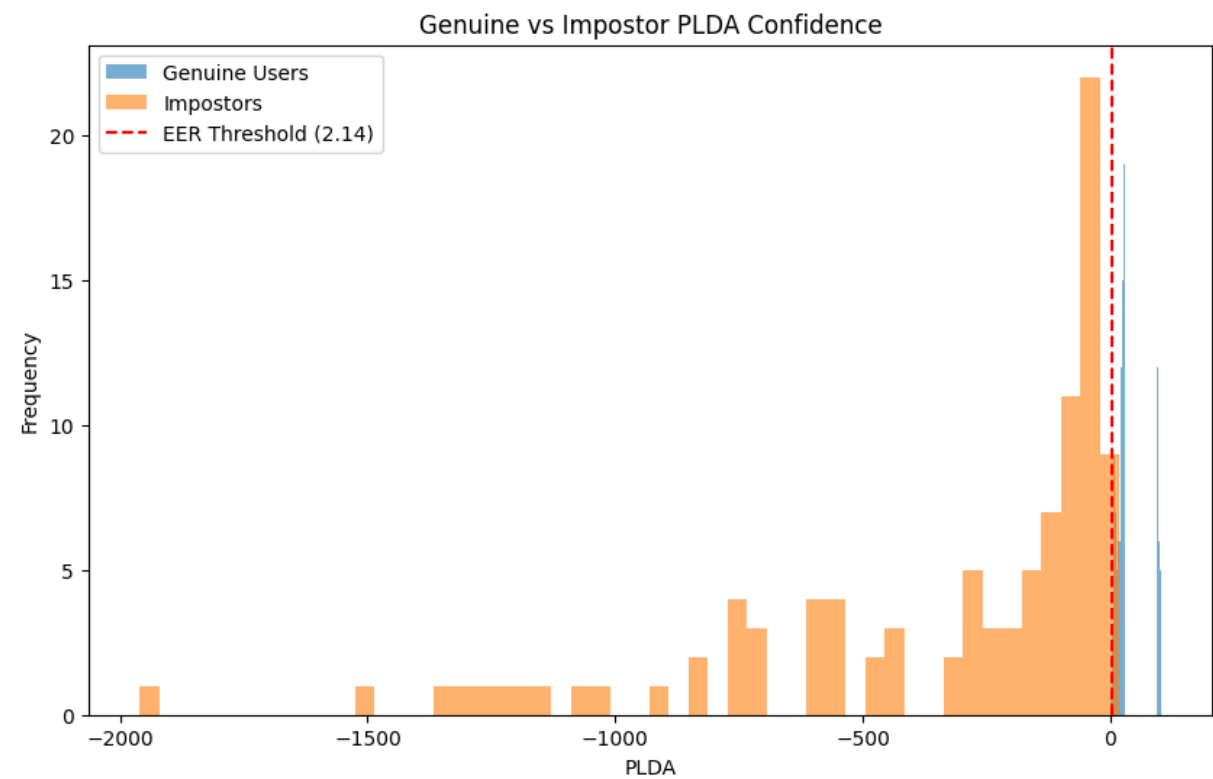
<i>Feature Extraction</i>		<i>Speaker Identification</i>	<i>Metriche</i>		
Features used	Frame Analysis	Model Used	Acc	F1	Precision
25 MFCCs	25ms frame length 10ms frame hop	DNN	1.0	1.0	1.0
25 MFCCs	25ms frame length 10ms frame hop	DNN	1.0	1.0	1.0
25 MFCCs	25ms frame length 10ms frame hop	RNN	0.97	0.96	0.97
25 MFCCs	25ms frame length 10ms frame hop	CNN+LSTM	0.99	0.99	0.99
20 MFCCs	25ms frame length 10ms frame hop	I-DNN	0.98	0.98	0.98
24 MFCCs, 24 Delta-MFCCs VAD 30% energy	25ms frame length 10ms frame hop	TDNN	0.94	0.94	0.95

Table 5.2: Speaker Identification

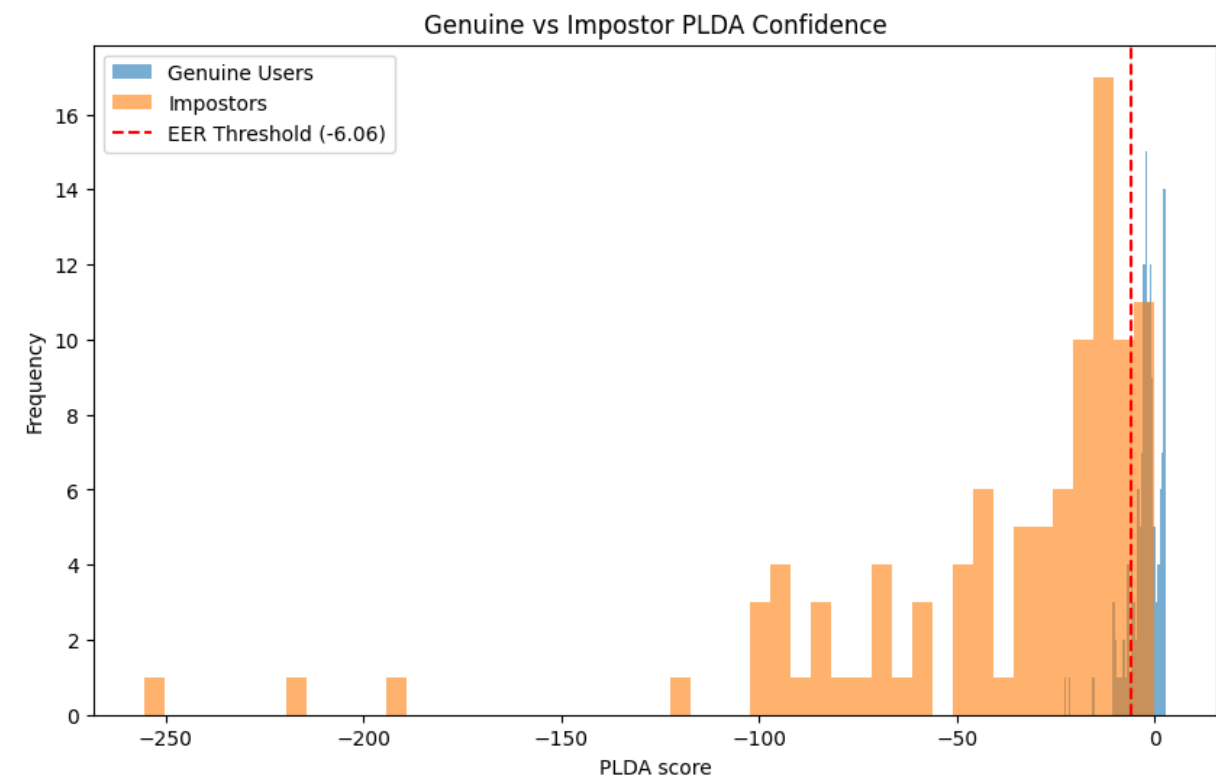
<i>Feature Extraction</i>		<i>Speaker Identification</i>	<i>Speaker Verification</i>		<i>Metrics</i>
Features used	Frame Analysis	Model Used	Embedding	Backend	EER
25 MFCCs	25ms frame length 10ms frame hop	DNN	Bottle-neck layer	GMM-UBM	0.03
25 MFCCs	25ms frame length 10ms frame hop	DNN	Last Hidden layer	GMM-UBM	0.13
25 MFCCs	25ms frame length 10ms frame hop	RNN	Mean Hidden Layers	GMM-UBM	0.12
25 MFCCs	25ms frame length 10ms frame hop	CNN+LSTM	Mean Hidden Layers	GMM-UBM	0.15
20 MFCCs	25ms frame length 10ms frame hop	I-DNN	I-embedding	GMM-UBM	0.16
24 MFCCs, 24 Delta-MFCCs VAD 30% energy	25ms frame length 10ms frame hop	TDNN	X-embedding	GMM-UBM	0.48

Table 5.3: Speaker Verification

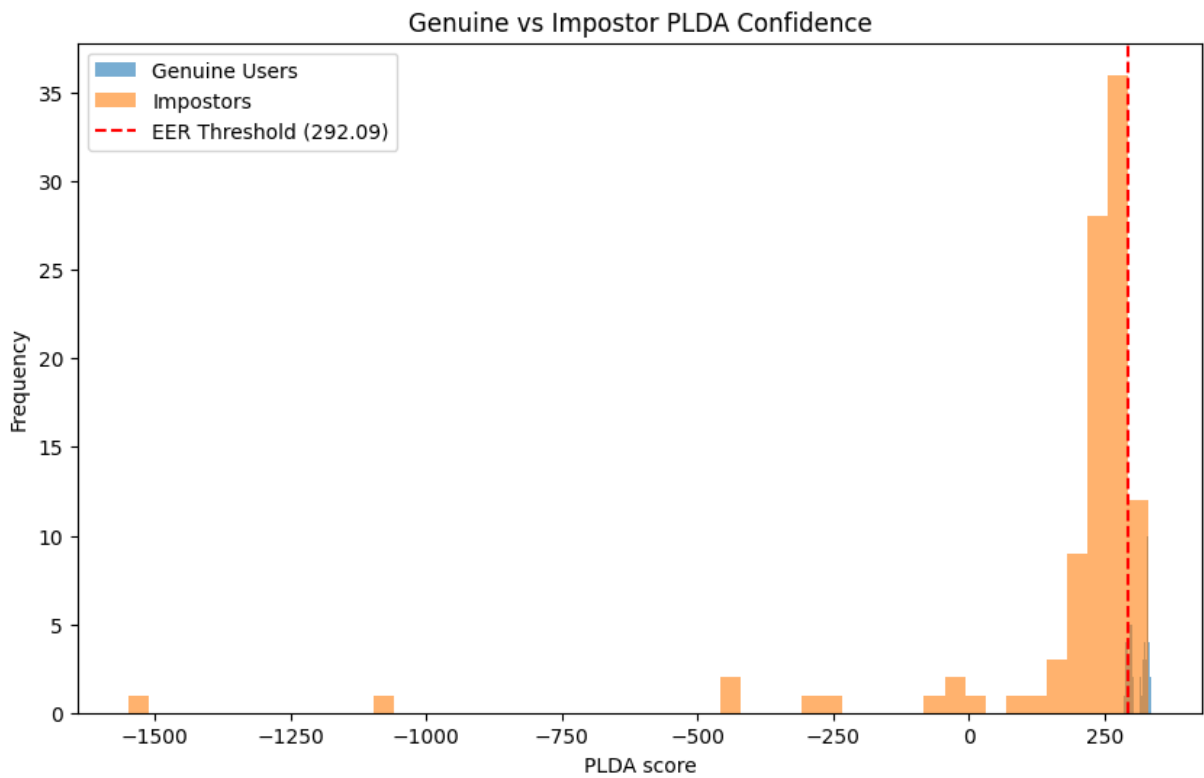
Architettura: DNN-1



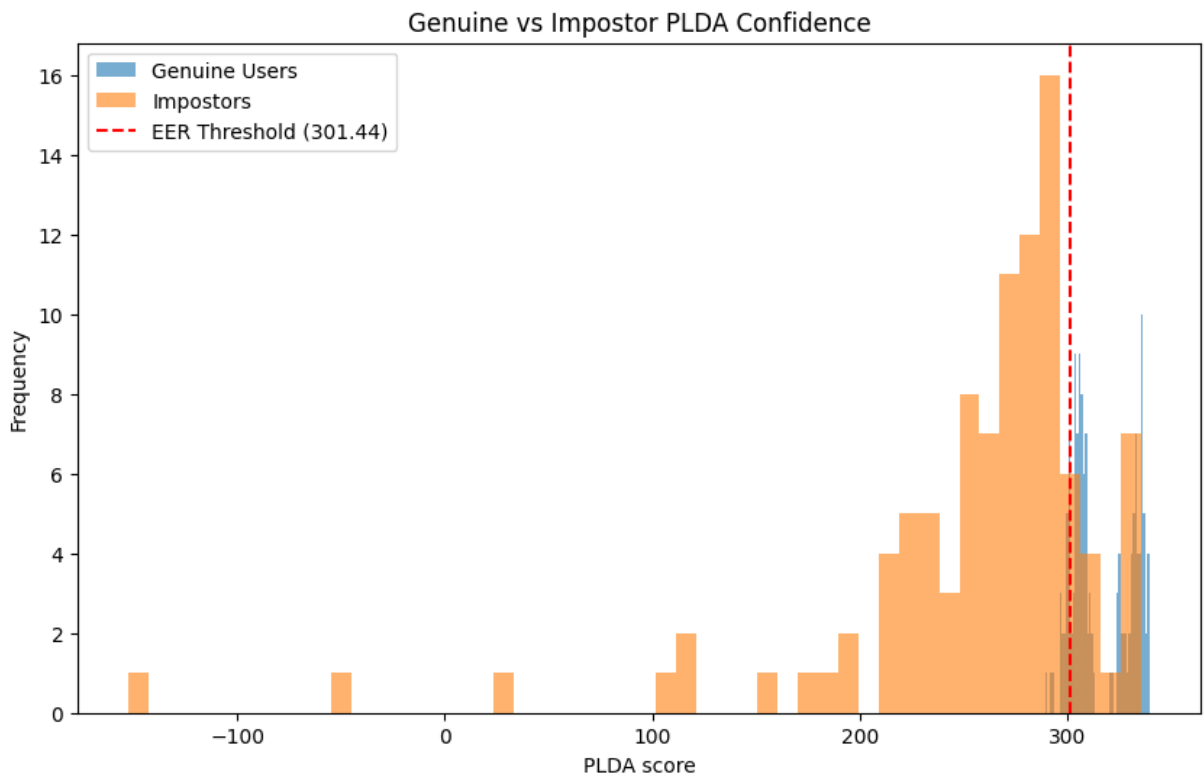
Architettura: DNN-2



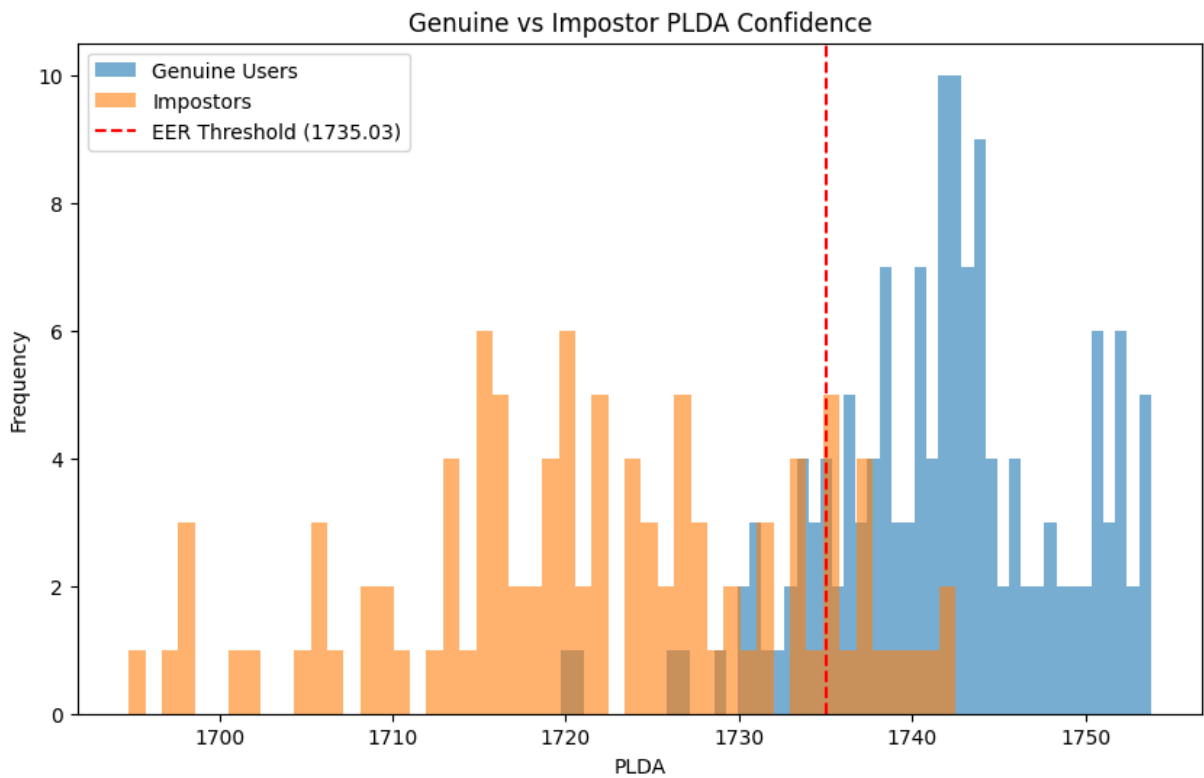
Architettura: RNN



Architettura: CNN+LSTM



Architettura: I-DNN



Architettura: X-DNN

