

Esame di Fondamenti di Visione Artificiale e Biometria: EmotionGait

Afeltra Angelo, Strianese Davide Benedetto

Sommario

Le emozioni svolgono un ruolo importante nella nostra vita quotidiana, infatti influenzano il nostro benessere, le relazioni interpersonali, le decisioni che prendiamo e il modo in cui reagiamo alle situazioni in cui ci troviamo. Le emozioni possono variare molto e comprendere una vasta gamma di stati come la gioia, la tristezza, la rabbia, e molte altre.

L'obiettivo di questo lavoro è quello di capire se sia possibile individuare l'emozione provata da una persona dalla sua camminata. Il progetto, analizzando l'andatura di una persona, estrae alcune caratteristiche e in base a queste cerca di capire l'emozione che sta provando.

La prima parte dell'elaborato sarà dedicata ad un'introduzione relativa alle tematiche trattate e a diversi lavori presenti in letteratura che sono stati utili al progetto svolto. Saranno poi illustrate nei dettagli le strategie implementative adottate con i vari modelli usati nella fase di apprendimento. Inoltre, verranno mostrati e analizzati i risultati ottenuti. Infine, verranno esposte le conclusioni.

1. Introduzione

Le emozioni giocano un ruolo importante nelle nostre vite, definiscono le nostre esperienze, modellano il modo in cui vediamo il mondo e definiscono il modo in cui interagiamo con gli altri. Percepire le emozioni di una persona con cui stiamo interagendo ci aiuta a capire il suo comportamento e il suo stato d'animo e, attraverso queste informazioni, riusciamo a determinare il modo migliore per porci nei suoi confronti.

Infatti, a seconda dell'emozione che la persona trasmette ci comportiamo in maniera diversa, se è triste cerchiamo di confortarla, se è arrabbiata cerchiamo invece di calmarla e così via. Inoltre anche le emozioni di individui sconosciuti possono influenzare il nostro comportamento, infatti se per strada vediamo un pedone terrorizzato o felice cambia anche il nostro modo di agire.

Data l'importanza ricoperta dalle emozioni, il riconoscimento automatico di queste ultime è un problema critico in molti campi, come i giochi, l'intrattenimento, e la sicurezza. Solitamente gli esseri umani percepiscono le emozioni sia tramite il modo di parlare sia tramite il linguaggio del corpo. Negli ultimi tempi si sta cercando di simulare questo comportamento anche con dei dispositivi che grazie all'aiuto di intelligenze artificiali cercano di comprendere le emozioni analizzando il modo di parlare e di agire di una persona.

L'obiettivo di questo lavoro è quello di capire se sia possibile percepire l'emozione provata da una persona analizzando la sua camminata.

2. Lavori correlati

In letteratura sono presenti diversi lavori che trattano lo studio dell'emotion gait. Tra questi lavori ci è tornato molto utile il paper [5] il quale tratta come creare un modello di machine learning per identificare l'emozione di un soggetto analizzando la sua camminata. Questo paper, avendo il nostro stesso obiettivo, è stato usato come punto di partenza per la creazione del nostro progetto e per capire

quale approccio avere con questo problema. Altri paper che sono tornati utili sono stati [3] e [1], utili per capire quali sono le feature più rilevanti da estrarre dalle pose. Infine per comprendere al meglio l'uso della libreria *MediaPipe* [4] è stato studiato l'articolo [2] dove viene presentata una pipeline di rete neurale utile all'estrazione di landmark 3D e allo studio della posa.

3. Dataset

In questo capitolo verrà descritto il dataset utilizzato per lo svolgimento del progetto. Il dataset che ci è stato fornito è composto da due parti, la prima è una raccolta di video di persone che camminano e la seconda un file csv.

Video Una parte del dataset è composta da 60 video. Ogni video mostra una persona che cammina frontalmente per pochi secondi. I video presentano le seguenti particolarità che hanno portato a dei problemi durante la fase di pre-processing:

- loop: in ogni video lo stesso ciclo di gait viene riprodotto più volte avendo così una serie di loop. Questo è stato un problema nella fase di pre-processing in quanto è stato necessario estrarre un singolo loop;
- volto censurato: in ogni video la persona ripresa ha il volto censurato. Questo ha portato dei problemi quando bisognava estrarre le pose, infatti il più delle volte risultavano al contrario perché, a causa della mancanza del volto, la libreria non riusciva a capire se la persona fosse orientata frontalmente o di spalle.

Nella figura 1 vengono riportati degli screen presi da alcuni video del dataset.



Figura 1: Alcuni screen provenienti dai video

CSV La seconda parte comprende un file csv strutturato nel seguente modo:

- le colonne rappresentano il video della camminata di una persona e per ogni video sono previste 4 domande:

1. La persona nel video è felice?
2. La persona nel video è arrabbiata?
3. La persona nel video è triste?
4. La persona nel video è neutrale?

Quindi ci sono 4 colonne per ogni video per un totale di 240.

- Le righe rappresentano i partecipanti a cui sono state poste le domande. Nel caso di risposta il partecipante da un voto in una scala da 1 a 5 in cui i valori hanno i seguenti significati:

- 1 - fortemente in disaccordo
- 2 - un po' in disaccordo
- 3 - ne in accordo ne in disaccordo
- 4 - un po' in accordo
- 5 - fortemente in accordo

Nella figura 2 viene riportato un esempio di record in cui il partecipante 1 ha espresso dei voti per il video VID_RGB_062.

GaitID	VID_RGB_062	Question	The person in the video appears ... - Happy	VID_RGB_062	Question	The person in the video appears ... - Angry	VID_RGB_062	Question	VID_RGB_062	Question	VID_RGB_062
Participant 1			2			2			2		3

Figura 2: Esempio di un record del file csv

4. Strategie implementative

In questo capitolo verrà mostrata la strategia implementativa adottata per lo svolgimento del progetto. Per la realizzazione del lavoro si è ricorso ad un approccio che prevede le seguenti fasi:

1. Estrazione statistiche dal file csv e labelling dei video;
2. Isolamento di un singolo loop dai video;

3. Estrazione e filtraggio delle pose;
4. Allineamento delle pose;
5. Estrazione delle feature;
6. Applicazione dei principali modelli di machine learning;
7. Applicazione di deep;
8. Analisi dei risultati.

4.1. Estrazione statistiche dal file csv e labelling dei video

Nella prima fase sono state raccolte alcune statistiche sulle votazioni espresse dai partecipanti, in particolare è stata calcolata la media, la moda e la deviazione standard per ogni emozione. In figura 3 vengono riportare le proporzioni del dataset, andando ad utilizzare l'emozione con la media più alta come etichetta.

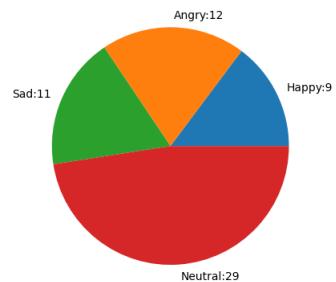


Figura 3: Media emozioni

In figura 4 invece vengono riportare le proporzioni del dataset di train utilizzando l'emozione con la moda più alta come etichetta.

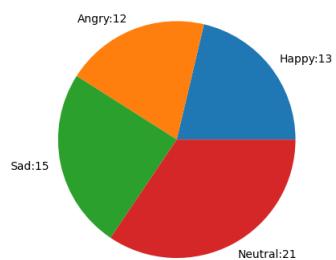


Figura 4: Moda emozioni

Da come si può notare utilizzando la media per generare le etichette il dataset risulta essere sbilanciato a favore dell'emozione neutral, infatti circa la metà dei video (29/61) risultano essere neutral. Andando ad osservare quella che è la deviazione standard per tale emozione, essa risulta essere abbastanza alta, pertanto si è deciso di utilizzare la moda per la generazione delle etichette.

4.2. Isolamento di un singolo loop dai video

Dopo un primo studio sul file csv si è passati ad analizzare i video. La prima problematica riscontrata è stata quella della presenza di più loop, lo stesso ciclo di gait viene ripetuto più volte all'interno del video, quindi bisognava isolare un singolo ciclo di gait da ogni video.

Per eseguire questa estrazione ci si è basati sulla similarità tra i frame (non è possibile utilizzare l'uguaglianza in quanto è presente del rumore), procedendo nel seguente modo:

- Selezione del primo frame
- Calcolo della similarità tra il primo frame e i successivi (primo col secondo, primo col terzo...)
- Confronto tra i valori di similarità ottenuti. Si ottiene una curva che ha un andamento crescente fino ad n frame per poi subire un calo brusco sul frame n+1. L'andamento crescente indica che la similarità di tali frame con il primo va a diminuire, mentre il punto di picco indica che quel frame ritorna ad essere molto simile al primo frame, pertanto siamo su un nuovo ciclo di gait.

In figura 5 è riportato un grafico contenente i valori di similarità. Come si può notare nel momento in cui si raggiunge un picco si ha l'inizio di un nuovo gait.

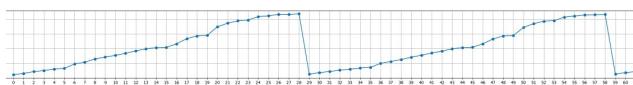


Figura 5: Grafico valori di similarità tra i frame

4.3. Estrazione e filtraggio delle pose

Nella terza fase vengono estratte e filtrate le pose frame per frame. Per l'acquisizione della posa è stata utilizzata la libreria MediaPipe che oltre ad effettuare l'estrazione della posa fissa dei landmark che coincidono con le parti principali del corpo. Nella figura 6 vengono raffigurati i landmark estratti dalla libreria.

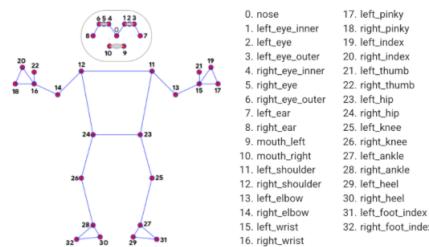


Figura 6: Landmark di Mediapipe

Nello studio della posa si è ritenuto opportuno non considerare i seguenti landmark:

- Landmark del volto: i punti da 0 a 10 non vengono presi in considerazione in quanto nei video forniti i volti sono censurati;
- Landmark delle mani: i punti da 15 a 22 non vengono considerati a causa di una scarsa qualità dei video che apporta delle problematiche alla libreria nell'individuare questi punti;
- Landmark dei piedi: per lo stesso motivo del punto precedente i landmark dei piedi che vanno da 27 a 32 vengono tralasciati.

Dopo aver acquisito le pose ci si è resi conto che alcune di esse non venivano individuate in maniera corretta per i seguenti motivi:

- Posa tracciata non corretta: come si può vedere dalla figura 7 la posa tracciata non corrisponde con la posa della persona presente nel video



Figura 7: Posa Clippata

- Orientamento della posa invertito: andando a controllare le coordinate dei landmark estratti, su alcuni frame la parte destra e la parte sinistra risultano invertire, questo indica che la libreria ha rilevato un cambio di orientamento della persona. In figura 8 vengono riportate le coordinate di una posa frame per frame e si può notare come ci sia un'inversione di esse in alcuni frame.

Frame	Spalla Dx	Spalla Sx	Bacino Dx	Bacino Sx
0	0.422640	0.574540	0.439837	0.525403
1	1.0424470	0.574527	0.439704	0.525403
2	2.0423907	0.574593	0.439874	0.525472
3	3.0430507	0.579124	0.439874	0.525421
4	4.0573450	0.433016	0.490538	0.469209
5	5.0582642	0.423862	0.524748	0.443937
6	6.0589532	0.423308	0.528454	0.440855
7	7.0588675	0.420727	0.529734	0.437289
8	8.0583345	0.421033	0.529651	0.438483
9	9.0581968	0.419076	0.532844	0.436026
10	10.0578420	0.412803	0.536013	0.437208
11	11.0578988	0.407024	0.527470	0.443868
12	12.0412226	0.565992	0.485982	0.523277
13	13.0544811	0.498110	0.520483	0.453041
14	14.0407229	0.564403	0.456228	0.524344
15	15.0532594	0.407012	0.487454	0.502237
16	16.0395581	0.655979	0.438306	0.534683
17	17.0378632	0.656081	0.425888	0.539664
18	18.0375153	0.571585	0.424629	0.542361
19	19.0372038	0.570301	0.418810	0.544731
20	20.0369735	0.574337	0.413750	0.541699

Figura 8: Coordinate Landmark

Le pose clippate sono state filtrate partendo dalla posa del primo frame e verificando se le successive subiscono una brusca variazioni per le seguenti misure: larghezza spalle, larghezza bacino, distanza tra i gomiti e distanza tra polsi e bacino. Inoltre, come ulteriore criterio per il filtraggio viene verificato se le pose tracciate contengono le braccia incrociate.

Per le pose con un orientamento sbagliato, si è deciso di non rimuoverle in quanto la loro eliminazione avrebbe causato la perdita di molti frame. Tuttavia è stato possibile utilizzarle in quanto il problema dell'orientamento è relativo alla coordinata z e lavorando solo sulle coordinate x ed y è stato possibile utilizzare queste pose. In figura 9 viene riportato un esempio di frame nel quale è stata riscontrata questa problematica.

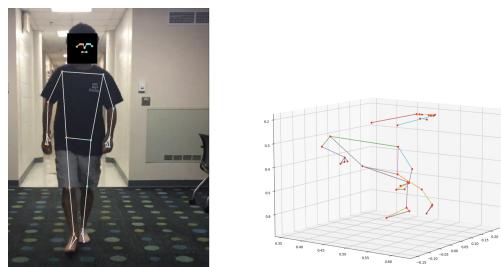


Figura 9: Posa con orientamento errato

4.4. Allineamento pose

La quarta fase comprende l'allineamento delle pose. Questa operazione consiste nell'ottenere lo stesso numero di pose per tutti i gait, in quanto sia precedentemente che successivamente alla fase di filtraggio ogni ciclo di gait possedeva un numero di pose differenti. Questa operazione è stata necessaria per favorire l'apprendimento del modello, in quanto andrà a lavorare su delle istanze che avranno la stessa dimensione. Per eseguire questa operazione è stato fissato un numero di pose, nel nostro caso 30 (corrisponde al frame rate dei video) e a seconda del numero di pose pulite di un video si procede nel seguente modo:

- minore di 30: per raggiungere il numero di pose prefissate vengono selezionate casualmente 2 pose consecutive con i rispettivi landmark. Successivamente tra di esse viene frapposta una nuova posa sintetica, ottenuta calcolando dei nuovi landmark come punti medi delle due pose scelte. Questa operazione viene iterata fino a quando non si raggiunge il numero di frame prefissato;
- maggiore di 30: vengono scartati dei frame casuali fatta eccezione per il primo e l'ultimo;
- pari a 30: non viene effettuata nessuna operazione.

4.5. Estrazione delle feature

L'ultima operazione, prima di passare all'identificazione di un modello, consiste nel selezionare le feature che dovranno essere usate per l'addestramento. Di seguito le feature scelte:

- Distanza tra la mano sinistra e la spalla sinistra
- Distanza tra la mano destra e la spalla destra
- Distanza tra la mano sinistra e il fianco sinistro
- Distanza tra la mano destra e il fianco destro
- Distanza tra il gomito sinistro e il fianco sinistro
- Distanza tra il gomito destro e il fianco destro
- Inclinazione delle spalle
- Area occupata dalla posa
- Distanza tra le caviglie
- Distanza tra la caviglia destra e l'anca destra
- Distanza tra la caviglia sinistra e l'anca sinistra

4.6. Applicazione dei principali modelli di machine learning

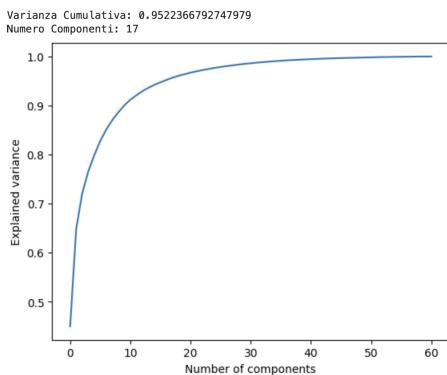
Nell'applicazione del machine learning sono stati eseguiti diversi esperimenti in cui sono state valutate le prestazioni dei principali modelli attraverso l'utilizzo di una cross validation con 5 fold. I modelli valutati sono stati:

Support Vector Machine, Decision Tree, Random Forest, XGB, Perceptron, Logistic Regression e K-nearest neighbors. La valutazione degli esperimenti è stata effettuata sulle seguenti pipeline:

1. Valutazione di una pipeline su dati grezzi;
2. Valutazione di una pipeline applicando la tecnica della PCA sui dati;
3. Valutazione di una pipeline applicando data augmentation tramite l'utilizzo di SMOTE;
4. Valutazione di una pipeline applicando data augmentation utilizzando tutti i cicli di gait presenti nel video;
5. Valutazione di una pipeline applicando PCA e SMOTE;
6. Valutazione di una pipeline applicando PCA e utilizzando tutti i cicli di gait;

Per quanto riguarda le pipeline in cui viene utilizzata la PCA, grazie a tale tecnica, si è riusciti a ridurre la dimensionalità del dato a 17 feature mantenendo il 95% di varianza cumulativa. In figura 10 viene riportato il grafico rappresentante l'utilizzo di PCA.

Nella pipeline in cui viene effettuato il processo di data augmentation tramite SMOTE va detto che questo viene eseguito solamente sul train set dopo aver effettuato lo split train-validation. In questo modo le prestazioni del modello non vengono stimate su dati sintetici. Invece nell'altra pipeline in cui viene fatto data augmentation utilizzando tutti i cicli di gait presenti in un video, lo split train-validation viene effettuato in modo tale che tutti i cicli di gait di uno stesso video appartengono ad uno dei due set.

**Figura 10:** PCA

Infine, per andare a stimare in maniera più realistica le prestazioni di un modello, nel validation set verrà mantenuto un solo ciclo di gai per ogni video.

I risultati di tali esperimenti verranno mostrati nel capitolo 5.

4.7. Applicazione di deep learning

Nell'applicazione del deep learning sono stati eseguiti diversi esperimenti per identificare i migliori iperparametri da utilizzare nella costruzione e nell'addestramento del modello. Di seguito è riportata la lista degli iperparametri testati:

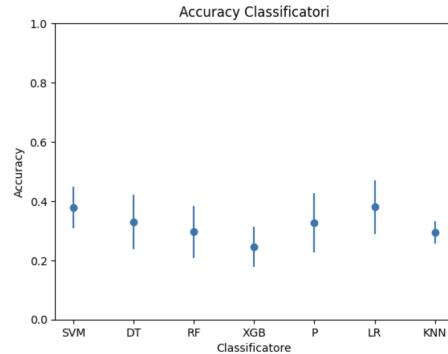
- Numero Neuroni: 4,8,16,32,64,128
- Numero Layer Nascosi: 0,1,2,3
- Ottimizzatore: SGD, Adam, AdamW, Adadelta
- Epoche: 100,200,500,1000,2000,3000

Il modello che ha riportato i risultati migliori è composto da 3 strati di layer: lo strato iniziale con 32 neuroni, uno strato nascosto con 16 neuroni e uno strato finale con 4 neuroni. Inoltre tra tali strati è stato aggiunto un Dropout con un rate di 0.5. Per quanto riguarda l'addestramento del modello il miglior ottimizzatore è risultato essere AdamW il quale impiega 2000 epoch per ottenere il miglior risultato. I risultati di tale modello verranno mostrati nel capitolo 5.

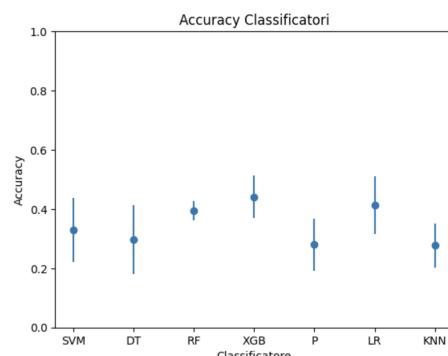
5. Risultati ottenuti

In questo capitolo verranno illustrati i risultati ottenuti sia con i modelli di machine learning sia con quelli di deep learning.

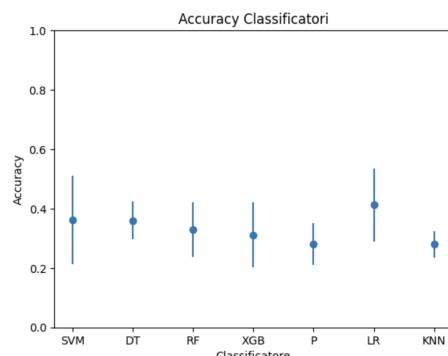
Esperimento 1: Pipeline senza pre-processing

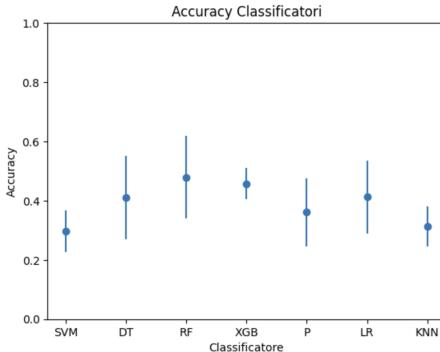
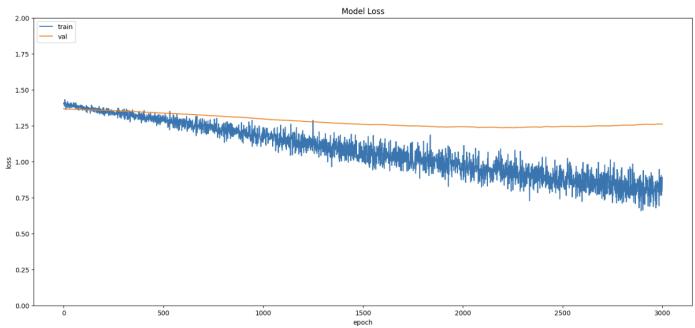
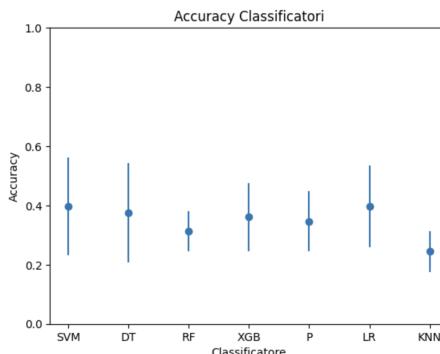
**Figura 11:** Risultati esperimento 1

Esperimento 2: Pipeline con PCA

**Figura 12:** Risultati esperimento 2

Esperimento 3: Data Agumentation con SMOTE

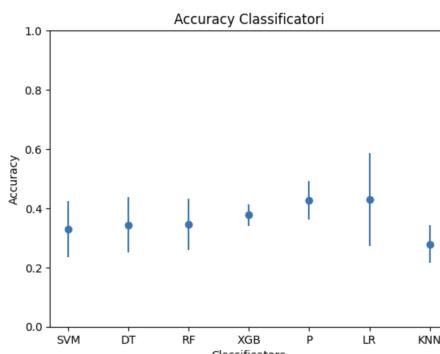
**Figura 13:** Risultati esperimento 3

Esperimento 4: PCA & SMOTE**Figura 14:** Risultati esperimento 4**Modello Deep Learning****Figura 17:** Model Loss**Esperimento 5: Data Augmentation riutilizzando tutti presenti nel video****Figura 15:** Risultati esperimento 5

	precision	recall	f1-score	support
0	0.89	0.80	0.84	10
1	0.82	1.00	0.90	9
2	0.89	0.67	0.76	12
3	0.79	0.88	0.83	17
accuracy			0.83	48
macro avg	0.85	0.84	0.83	48
weighted avg	0.84	0.83	0.83	48

Figura 18: Train performance

	precision	recall	f1-score	support
0	0.50	0.33	0.40	3
1	0.50	0.33	0.40	3
2	0.67	0.67	0.67	3
3	0.50	0.75	0.60	4
accuracy			0.54	13
macro avg	0.54	0.52	0.52	13
weighted avg	0.54	0.54	0.52	13

Figura 19: Val performance**Esperimento 6: PCA & Data Augmentation riutilizzando tutti presenti nel video****Figura 16:** Risultati esperimento 6**6. Analisi dei risultati**

Per il primo esperimento, [11], i risultati migliori si hanno con SVM e LogisticRegression che raggiungono un accuracy del 38%. Tali performance vengono migliorate nell'esperimento 2, [12], grazie all'aggiunta di PCA e XGB raggiungendo così un'accuracy del 44%. Nell'esperimento 3, 13, è stato effettuato data augmentation tramite SMOTE ottenendo un leggero miglioramento con accuracy pari al 41%, dove il risultato varia tra il 29% e 35%. Nell'esperimento 4, 14, è stato sempre effettuato data augmentation tramite SMOTE ma in più lo si è combinato con la PCA. In questo caso il Random Forest raggiunge prestazioni pari al 48% di accuracy, con un risultato che varia tra il 35% e il 60%. Da notare che l'XGB raggiunge un'accuracy inferiore pari al 46%, ma ha una variazione del risultato migliore, quindi può essere considerato migliore del Random Forest.

Negli esperimenti 5 e 6, 15 e 16, è stato effettuato nuovamente data augmentation considerando però tutti i cicli di gait nel singolo video. Anche in questo caso si è utilizzata la combinazione con la PCA per l'esperimento 6.

Nei due esperimenti il risultato migliore si ha con il Logistic Regression, raggiungendo rispettivamente un'accuracy del 39% e del 42%. Da notare che con questo tipo di data augmentation si ha che tutti i classificatori hanno una variazione consistente pari al +/- 10%.

Il grafico in figura 20 mostra il variare delle prestazioni in base ai vari esperimenti condotti:

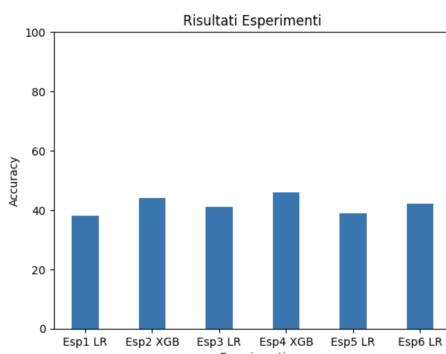


Figura 20: Summary Esperimenti

Per ottenere dei risultati utilizzando le reti neurali, sono stati testati diversi iperparametri fino a raggiungere un'accuracy pari al 54%. Anche se sembra un buon risultato, come riportato nel grafico 17, bisogna notare che la validation loss risulta avere un andamento quasi costante anche se è presente un minimo di ottimizzazione. Molto probabilmente, tale condizione è causata dalla mancanza di dati. Infatti, le reti neurali lavorano al meglio quando hanno a che fare con una grande mole di dati.

Analizzando i risultati ottenuti con i modelli del deep learning sembrerebbe che ci sia un netto miglioramento. Tuttavia non è così, infatti andando a valutare le prestazioni del modello tramite una cross-validation a 5 fold la media si aggira attorno al 45%. In conclusione non si ha nessun miglioramento rispetto ai modelli di machine learning.

7. Conclusioni

L'obiettivo di questo lavoro è stato quello di capire se fosse possibile ottenere l'emozione da un soggetto analizzando solo la sua camminata. Nella prima fase dell'elaborato sono stati illustrati i lavori già presenti in letteratura da cui si è preso spunto. Successivamente sono state mostrate le strategie implementative scelte ed i vari risultati ottenuti in termini di accuracy mostrando i grafici di loss e accuracy relativi alla fase di addestramento del modello. I risultati ottenuti sono stati, infine, analizzati ponendo attenzione anche su ciò che non è andato bene durante lo sviluppo. In conclusione, gli esperimenti hanno mostrato che è possibile rilevare l'emozione di una persona analizzando la sua camminata, validando la rete neurale costruita sul test-set il nostro approccio ha raggiunto un'accuracy del 40%.

	precision	recall	f1-score	support
0	1.00	0.25	0.40	4
1	0.00	0.00	0.00	3
2	0.33	0.33	0.33	3
3	0.36	0.80	0.50	5
accuracy			0.40	15
macro avg	0.42	0.35	0.31	15
weighted avg	0.45	0.40	0.34	15

Figura 21: Test Validation Deep Model

Tuttavia pensiamo che andando a migliorare la qualità del pre-processing e avendo a disposizione più dati, tali performance potranno essere migliorate. Lavori futuri potrebbero includere le seguenti opzioni:

- utilizzare una libreria per l'estrazione delle pose in cui la presenza del volto non è rilevante, in modo da avere pose 3D più accurate;
- utilizzare dei video in cui il volto del soggetto non sia censurato;
- acquisire più dati o in alternativa utilizzare una GAN per la generazione di pose sintetiche

Riferimenti bibliografici

- [1] Arthur Crenn et al. «Body expression recognition from animated 3D skeleton». In: *2016 International Conference on 3D Imaging (IC3D)*. IEEE. 2016, pp. 1–7.
- [2] Ivan Grishchenko et al. «BlazePose GHUM Holistic: Real-time 3D Human Landmarks and Pose Estimation». In: *arXiv preprint arXiv:2206.11678* (2022).
- [3] Michelle Karg, Kolja Kühnlenz e Martin Buss. «Recognition of affect based on gait patterns». In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40.4 (2010), pp. 1050–1061.
- [4] Mediapipe. URL: <https://google.github.io/mediapipe/>.
- [5] Tanmay Randhavane et al. «Identifying emotions from walking using affective and deep features». In: *arXiv preprint arXiv:1906.11884* (2019).