

Valutazione dell'estetica di immagini di cibo

Relatore: *Prof. Paolo Napoletano*

Correlatore: *Prof. Gianluigi Ciocca*

Relazione della prova finale di:

Sofia Damaso

Matricola 845189

Anno Accademico 2020-2021

*Ai miei genitori per avermi sostenuta fin da piccola di fronte
a tutte le piccole e grandi difficoltà.*

*A Gabriele con il quale ho un rapporto importante,
basato sulla stima e sul supporto reciproco, e che mi ha aiutato
con alcune delle fotografie presenti in questa relazione.
Proprio da questo rapporto è nata l'idea di mettermi alla prova
con un progetto legato alle immagini e al loro studio.*

*Alle mie amiche che mi hanno vista crescere,
compiendo le scelte più grandi della mia vita,
e alla mia amica Ilaria, una persona speciale sempre
pronta a consigliarmi e a starmi accanto ogni giorno.*

*Ai miei compagni di corso che in questi anni sono
stati amici con cui passare le giornate di studio.
In particolare a Edoardo e Daniel che sono sempre
stati pronti ad aiutarmi nei momenti più difficili
di questo percorso e con i quali ho condiviso molto.*

*Ai professori che mi hanno aiutata e formata
per affrontare questo progetto nel migliore dei modi,
fornendomi sempre nuovi stimoli e spunti di riflessione.*

*A me stessa, per non essermi mai persa d'animo
nei momenti di difficoltà perché sapevo di avere le
capacità per raggiungere i miei obiettivi.*

Audentes Fortuna iuvat.

Indice

Introduzione	1
1 Il tema dell'estetica	3
1.1 L'importanza dell'estetica	3
1.2 Estetica e cibi	7
1.3 Outline	9
2 Gourmet Photography Dataset ed esperimenti	10
2.1 Contenuto e sviluppo di GPD	10
2.2 Test e risultati	14
3 Dataset proposto	17
3.1 Motivazione e composizione	17
3.2 Assegnamento delle groundtruth	18
4 Metodologie per la classificazione	29
4.1 Feature hand-crafted	29
4.2 Un'ulteriore suddivisione delle immagini	29
4.3 Feature estratte da una rete neurale	30
4.4 Uso di una rete neurale per l'intera classificazione	31
4.4.1 Prima implementazione	31
4.4.2 Early Stopping	31
4.5 Valutazione del dataset proposto	32
4.6 Grad-CAM per capire le predizioni	33
5 Risultati ottenuti	36
5.1 Risultati ottenuti sul dataset GPD	36
5.1.1 Feature hand-crafted	36
5.1.2 Feature estratte da una rete neurale	37
5.1.3 Uso delle reti neurali per l'intera classificazione	37
5.2 Risultati ottenuti sul dataset proposto	38
5.2.1 Analisi delle groundtruth	38

5.2.2	Uso di una rete neurale adattata	39
5.2.3	Analisi delle Grad-CAM	41
6	Conclusioni e sviluppi futuri	44
Riferimenti		46
Bibliografia	46	
Siti	48	

Elenco delle figure

1.1	Esempi di immagini che non vengono considerate esteticamente belle poiché sono sottoesposte, sovraesposte o sfocate	4
1.2	Esempi di immagini che vengono considerate esteticamente belle poiché lo scatto è equilibrato e ben bilanciato	5
1.3	Alcune delle 265 milioni di immagini che fanno parte dell'hashtag Foodporn su Instagram	8
2.1	Esempio di alcune immagini del GPD che a cui è associata la label negativa	11
2.2	Esempio di alcune immagini del GPD che a cui è associata la label positiva	12
2.3	Schema tratto dal lavoro originale [19] che illustra la pipeline della valutazione delle immagini del GPD da parte di 57 lavoratori del AMT (Amazon's Mechanical Turk)	13
3.1	Esempio di una delle domande del questionario che è servito per far valutare a 41 utenti le immagini del dataset proposto e ottenere le label di groundtruth	18
3.2	Immagini del dataset a cui è stata assegnata una groundtruth negativa tramite i voti degli utenti	20
3.3	Immagini del dataset a cui è stata assegnata una groundtruth positiva tramite i voti degli utenti	25
3.4	Esempio di una delle domande del questionario che è servito per far valutare a 11 utenti le immagini del dataset proposto e a motivare il perché di tale valutazione al fine di comprendere su che aspetti dell'immagine si fossero focalizzati	26
3.5	Esempio di una immagine che ha ottenuto solo voti negativi nel secondo questionario, principalmente a causa del fatto che il cibo non ha una composizione accurata, è già stato mangiato e la fotografia è leggermente sfocata	27

3.6	Esempio di una immagine che ha ottenuto solo voti positivi nel secondo questionario, gli utenti si sono concentrati principalmente sui colori molto vivaci che formano un bel contrasto cromatico con il tavolo e sulla composizione accurata dei poke	28
4.1	Esempio di una immagine a cui è stata applicata la tecnica Grad-CAM focalizzandosi su due differenti classi di oggetti, in questo caso prima sulla classe "Gatto" e poi sulla classe "Cane", il quale è stato tratto dal lavoro originale sulla tecnica Grad-CAM [17]	33
4.2	Esempi di maschere binarie dei cibi, le quali mostrano in bianco il cibo e in nero il background	34
5.1	Esempi di immagini a cui è stata applicata la tecnica Grad-CAM dove la rete ha predetto correttamente una label negativa	39
5.2	Esempi di immagini a cui è stata applicata la tecnica Grad-CAM dove la rete ha predetto correttamente una label positiva	40
5.3	Esempio di una immagine per la quale la rete ha predetto la label negativa mentre la groundtruth era positiva, per cui essa rientra tra gli errori e in particolare tra i falsi negativi. A destra viene riportata anche la rispettiva Grad-CAM, dalla quale si nota che le aree che la rete ha considerato più significative per la predizione della label non sono quelle del cibo, bensì appartengono al piatto e allo sfondo	40
5.4	Esempio di una immagine per la quale la rete ha predetto la label positiva mentre la groundtruth era negativa, per cui essa rientra tra gli errori e in particolare tra i falsi positivi. A destra viene riportata anche la rispettiva Grad-CAM, dalla quale si nota che le aree che la rete ha considerato più significative per la predizione della label sono effettivamente quelle del cibo e non altre parti dell'immagine	41
5.5	Grafico che rappresenta sull'asse delle x l'indicatore di concentrazione per ognuna delle immagini del dataset proposto, mentre sull'asse delle y rappresenta le label predette, in particolare il valore -1 indica la label negativa mentre il valore 1 indica la label positiva	42

Elenco delle tabelle

1.1	Elenco di alcuni dataset esistenti relativi allo studio dell'estetica in diversi ambiti e scopi	6
2.1	Livelli di accuratezza e risultati raggiunti nel lavoro originale [19] sul training set e sul test set utilizzando diversi approcci: combinazioni del classificatore SVM con diverse feature, reti neurali e reti neurali combinate con metodi di regolarizzazione	15
5.1	Livelli di accuratezza ottenuti sul training set e sul test set utilizzando diverse combinazioni di feature e il classificatore SVM	36
5.2	Livelli di accuratezza ottenuti sul validation set e sul test set utilizzando il classificatore SVM con delle feature estratte da due diversi layer della rete ResNet-18. Nel primo caso le feature sono state estratte dal layer pool5, il quale si trova alla fine della rete, mentre nel secondo caso dal layer res3b_relu, il quale si trova a metà della rete	37
5.3	Livelli di accuratezza ottenuti sul validation set e sul test set utilizzando una procedura di Fine Tuning su una rete ResNet-18. Sono stati riportati anche i parametri utilizzati per inizializzare il numero di epoche, la dimensione del batch e il tasso di apprendimento iniziale della rete	38
5.4	Livelli di accuratezza ottenuti sul validation set e sul test set utilizzando una procedura di Fine Tuning su una rete ResNet-18 con Early Stopping dopo una o due epoche, a seconda di quale valore portasse a un maggior risparmio di tempo, anche se nel caso di stop anticipato dopo una sola epoca con Initial Learn Rate pari a 0.00001 non c'è stato alcun miglioramento in quanto sono state eseguite comunque tutte le epoche del training. Sono stati riportati anche i parametri utilizzati per inizializzare il numero di epoche, la dimensione del batch e il tasso di apprendimento iniziale della rete	38

Introduzione

L’obiettivo di questa esperienza di stage, svoltasi in remoto presso l’Ateneo, è stata l’analisi di un dataset di immagini relative a cibi e la loro classificazione estetica.

Il linguaggio utilizzato per lo sviluppo di codice è stato interamente MATLAB [12], poiché è stato il linguaggio con cui mi sono approcciata all’elaborazione delle immagini nell’omonimo corso e, per questo motivo, è stato più semplice per me prendere spunto da ciò che era stato visto a lezione e a laboratorio per approcciarmi al problema dell’estetica.

Il tema dell’estetica è molto particolare in quanto, come verrà approfondito nel Capitolo 1, questa caratteristica dipende fortemente da chi osserva ma può e potrà essere sfruttata per l’analisi automatica di fotografie in svariati ambiti e in svariate applicazioni. In questa relazione ci si è concentrati sul cibo poiché anche in questa categoria di immagini è molto utile uno studio a livello estetico, ad esempio per il marketing e la pubblicità di ristoranti, e sarebbe utile che questa analisi possa essere svolta in maniera automatica.

Il lavoro svolto durante l’esperienza di stage è stato articolato nelle seguenti parti:

1. **Implementazione dei descrittori hand-crafted.** È stato scritto il codice MATLAB per estrarre dal dataset iniziale delle feature semplici, ad esempio quelle relative a colore e texture, ed è stato utilizzato un classificatore per conoscere l’accuratezza di classificazione con questi descrittori.
2. **Implementazione dell’estrazione dei descrittori da una rete neurale.** È stata scelta una rete neurale da cui sono state estratte delle feature con un livello di astrazione maggiore rispetto alle precedenti e, usando lo stesso classificatore del punto precedente, è stato calcolato il livello di accuratezza raggiunto.
3. **Utilizzo di una rete neurale per l’intero task.** È stata usata una rete neurale per l’intera fase di classificazione. La rete in questione era preaddestrata per svolgere un altro compito, ma è stato eseguito un Fine Tuning, ovvero una modifica di determinati layer per poterla adattare all’analisi delle immagini in questione.

4. **Creazione di un nuovo dataset e raccolta dati.** È stato proposto un nuovo dataset di immagini grazie all'aiuto di un fotografo professionista, le quali sono state valutate da 41 utenti e successivamente da un sottointeressante di 11 di essi al fine di comprendere come una persona valuti l'estetica e su cosa si basi per questa valutazione.
5. **Analisi dei risultati ottenuti e degli errori.** Sono stati fatti diversi ragionamenti sui risultati ottenuti dalle due valutazioni degli utenti ed è stato utilizzato un metodo per visualizzare le parti dell'immagine più significative per la rete neurale durante la predizione della label, in modo tale da osservare eventuali concordanze tra le label assegnate dagli utenti e le predizioni automatiche.

L'ultima fase del lavoro, ovvero l'analisi dei risultati e di ciò che ha portato la rete neurale a compiere una determinata predizione, è stata la fase più utile e interessante in quanto ha permesso di comprendere a fondo il ragionamento alla base delle predizioni. Questo è molto significativo poiché trovare le motivazioni dietro a un avvenimento è stato alla base della scelta di questa esperienza di stage e, in maniera più estesa, alla base della scelta di questo percorso universitario.

1

Il tema dell'estetica

1.1 L'importanza dell'estetica

Si dice che la bellezza stia negli occhi di chi guarda, ma misurare l'estetica è un task complesso da svolgere automaticamente [8] attraverso l'intelligenza artificiale. L'idea di riuscire a quantificare la bellezza è molto stimolante in quanto può essere sfruttata in diversi ambiti, ad esempio sistemi biometrici oppure sistemi robotici che permettono di scattare fotografie automaticamente, scartando quelle considerate non ottimali oppure per poter effettuare un enhancement automatico.

Quest'ultima possibilità è di particolare rilevanza in quanto le immagini digitali sono di largo impiego per la medicina, la sicurezza, la comunicazione, l'intrattenimento e molto altro, quindi la possibilità di individuare automaticamente immagini non soddisfacenti rispetto a determinati criteri estetici può portare a implementare algoritmi che possano elaborare una versione migliorata dell'immagine di partenza, la quale può essere utile per analisi o elaborazioni successive. Alcuni esempi di immagini che sono considerate esteticamente belle e immagini che non sono considerate tali sono riportati rispettivamente in Figura 1.2 e Figura 1.1.

Un problema a cui fare fronte è che l'estetica è una caratteristica fortemente soggettiva e quindi ci saranno conflitti tra le valutazioni assegnate da diversi utenti, le quali porteranno a numerosi problemi quando si strutturano algoritmi di valutazione automatica delle immagini.

Come verrà sottolineato nel corso di tutto il lavoro, l'uso delle reti neurali gioca un ruolo determinante al fine di classificare l'estetica di una serie di immagini

poiché è possibile ottenere risultati migliori rispetto alle feature hand-crafted e per questo nello stato dell'arte sono fortemente sfruttante all'interno degli algoritmi per lo studio dell'estetica.

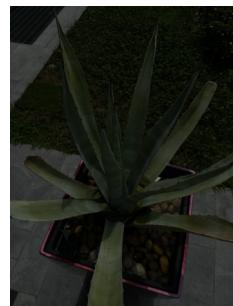


Figura 1.1: Esempi di immagini che non vengono considerate esteticamente belle poiché sono sottoesposte, sovraesposte o sfocate



Figura 1.2: Esempi di immagini che vengono considerate esteticamente belle poiché lo scatto è equilibrato e ben bilanciato

Di pari passo con l'aumento dell'utilizzo delle fotografie digitali cresce anche il numero di dataset di immagini per lo studio dell'estetica e la relativa dimensione, alcuni esempi sono riportati in Tabella 1.1.

Dataset	Amount	Domain
CUHK-PQ[11]	~17k	general image
AVA[15]	~250k	general image
AesCHN[22]	1k	Chinese handwriting
AutoTriage[2]	~16k	general image
AADB[9]	10k	general image
PCCD[3]	~4k	photo captioning
BlendPhotos[7]	1305	image blending
AesClothing[26]	-	clothing recommendation
GPD[19]	24k	food aesthetics

Tabella 1.1: Elenco di alcuni dataset esistenti relativi allo studio dell'estetica in diversi ambiti e scopi

In particolare, come è visibile nell'ultima riga della Tabella 1.1, il GPD (Gourmet Photography Dataset) [19] è incentrato sull'estetica dei cibi trattata come un problema di classificazione binaria, tema che verrà approfondito nel Capitolo 2.

Inoltre un dataset molto famoso è AVA (Aesthetic Visual Analysis) [15], il quale contiene 250000 immagini scattate da fotografi professionisti e contiene anche tre diversi tipi di annotazioni:

1. **Annotazioni relative all'estetica**, a ogni immagine è associato un punteggio che si basa sui voti degli utenti che hanno valutato le immagini.
2. **Annotazioni relative alla semantica**, a ogni immagine vengono associati uno o più tag tra i 66 disponibili.
3. **Annotazioni relative allo stile fotografico**, a ogni immagine viene associato uno stile tra 14 possibili, alcuni esempi sono High Dynamic Range, Regola dei terzi e Silhouette.

L'obiettivo di questo dataset, oltre ad essere uno dei dataset più grandi relativi allo studio dell'estetica, è anche quello di investigare la correlazione tra numerosità del dataset, qualità delle immagini utilizzate per il training e aumento della performance.

Nel caso del dataset CUHK-PQ [11] si hanno invece circa 17000 immagini, le quali sono sia immagini professionali di alta qualità che immagini di qualità inferiore scattate da utenti. Sono evidenti le differenze con il dataset AVA, sia per quanto riguarda la numerosità che per quanto riguarda la tipologia di immagini che compongono i due dataset. Un'ulteriore differenza è che in CUHK-PQ le immagini sono state divise manualmente in base al loro contenuto in 7 categorie:

1. **Animali**
2. **Piante**
3. **Architettura**
4. **Paesaggi**
5. **Oggetti statici**
6. **Persone**
7. **Notte**

Con l'incremento dell'utilizzo dei social network diventa molto interessante anche la tematica della selezione delle fotografie dopo che se ne sono scattate molte in serie, per questo è stato creato il dataset AutoTriage [2], il quale contiene 15545 fotografie organizzate in 5953 serie. Queste immagini provengono da album personali di utenti e non sono state modificate o selezionate precedentemente, altrimenti si perderebbe l'obiettivo finale di questo dataset.

È necessario sottolineare che la raccolta delle immagini per questo dataset è risultata complessa in quanto gli utenti generalmente prima di pubblicare delle fotografie sui social network le selezionano ed eventualmente le modificano con filtri o altre tecniche, sarebbe possibile ottenere immagini senza modifiche da siti che offrono archiviazione cloud ma a causa della privacy non è possibile l'accesso a terze parti.

1.2 Estetica e cibi

Molto spesso nella Filosofia la bellezza è stata legata alla bontà in diversi contesti ed è lampante quanto questa assunzione sia applicabile a casi in cui ci si trova davanti a un piatto oppure a una foto di esso, ad esempio consultando un menù o il sito web di un ristorante, infatti per l'uomo l'estetica è un aspetto importante quando si trova a valutare qualcosa.

In Psicologia si parla di questa tematica riferendosi ad essa con il nome di effetto alone, ovvero che ciò che appare bello agli occhi delle persone è implicitamente percepito come buono e gli vengono associate caratteristiche positive. Questo ragionamento viene spesso applicato quando si osserva la fotografia di una persona [4], ma anche in questo caso può essere applicato all'ambito delle fotografie di cibo infatti, spesso, quando un utente osserva una fotografia che considera esteticamente bella automaticamente pensa che quel cibo sia appetitoso. L'effetto alone mostra anche quanto la tematica dell'estetica sia attuale, nei social network le persone

vogliono apparire belle e ciò si riflette anche nelle fotografie di cibo, che vengono prese come riferimento dai ristoranti come se ci fosse una gara a chi crea il piatto più bello da vedere.

Nel corso degli anni l'interesse in questa tematica si è spostato dai filosofi e dagli psicologi alla comunità tecnica e scientifica [14], in particolare è significativa per la computer vision.

La scelta di concentrarsi sulle immagini di cibo nasce perché si tratta di un ambito sottosviluppato e su cui c'è ancora molto lavoro da svolgere, ma che è molto promettente soprattutto nel mondo del marketing e del commercio. Inoltre con l'avvento degli smartphone e dei social network un numero sempre maggiore di utenti ha iniziato a fotografare e pubblicare numerose foto di cibo, ad esempio come è visibile in Figura 1.3 l'hashtag Foodporn su Instagram ha un numero di post in costante crescita, ad oggi pari a 265 milioni.

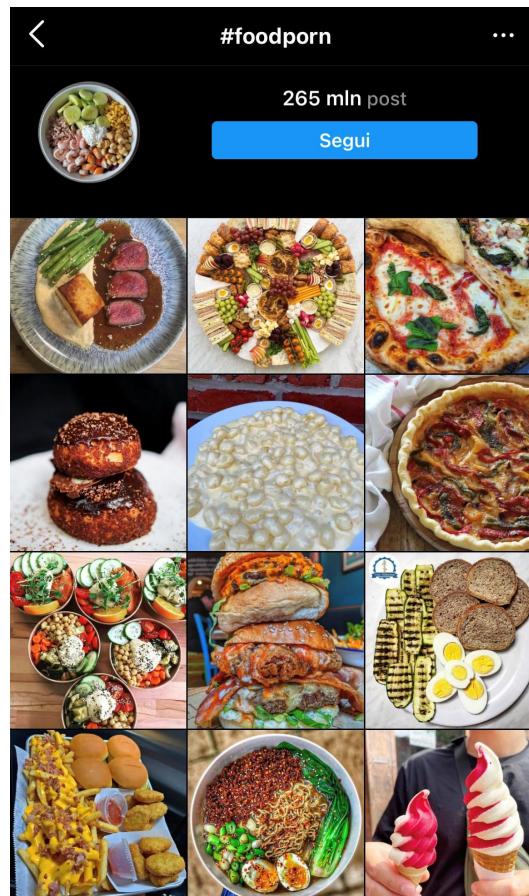


Figura 1.3: Alcune delle 265 milioni di immagini che fanno parte dell'hashtag Foodporn su Instagram

Una caratteristica comune delle immagini appartenenti a questo hashtag è che esse evocano nell'utente voglia di assaporare quei cibi, inoltre si è osservato che i cibi poco salutari e i dolci spesso diventano popolari sui social network poiché quelli salutari vengono percepiti dalle persone come meno gustosi, infatti nel 2016 si è visto che all'interno di #foodporn erano predominanti cibi non salutari [16].

Generalmente i cibi visibili sui social network o nelle pubblicità tendono anche a diventare dei modelli di estetica a cui i ristoranti cercano di avvicinarsi, cercando sempre più modi per presentare in maniera esteticamente bella i piatti. La presentazione di un piatto gioca un ruolo fondamentale quando si parla della percezione di un utente che lo osserva, in quanto in caso di presentazione del piatto molto articolata o artistica [13] gli utenti sono disposti a pagare un prezzo maggiore poiché considerano quel piatto esteticamente bello, molto elaborato e curato.

Il continuo interesse nella presentazione dei cibi scaturisce dal fatto che il mangiare comincia già con la vista, in particolare il gusto non è solo legato al sapore di un cibo ma dipende anche da tutti gli altri sensi [1]. L'aspetto di un cibo ha anche particolari effetti sul corpo umano: viene stimolata la salivazione [25], si attivano particolari aree del cervello come l'Amigdala [10] e si creano delle aspettative per quanto riguarda il gusto del cibo stesso.

Un esempio riguarda il colore [21] del cibo, in particolare se una bevanda è di colore verde si penserà che essa abbia un gusto più aspro rispetto a una bevanda di colore rosso. Inoltre è possibile che anche la forma di un cibo influenzi la nostra percezione [24], un esempio è legato al fatto che la dolcezza viene associata a cibi con forma tondeggiante, mentre forme squadrate o spigolose sono associate cibi salati, amari o aspri.

1.3 Outline

Questa relazione si concentra in primo luogo sulla descrizione dei due dataset utilizzati, ovvero il Gourmet Photography Dataset e un dataset creato appositamente, rispettivamente nel Capitolo 2 e nel Capitolo 3.

Successivamente nel Capitolo 4 verranno descritti i metodi utilizzati per estrarre le feature usate per le classificazione e nel Capitolo 5 i rispettivi risultati ottenuti. Inoltre sempre nel Capitolo 4 e nel Capitolo 5 verrà mostrato come visualizzare quali parti di immagine pesano maggiormente per una rete neurale che deve predire la classe di appartenenza di una immagine, dove la classe può essere positiva oppure negativa in base alla valutazione estetica dell'immagine stessa.

Il Capitolo 6, posto in chiusura della relazione, sarà dedicato alle conclusioni e ai possibili sviluppi futuri del lavoro svolto, individuando eventualmente possibili criticità e punti a favore.

2

Gourmet Photography Dataset ed esperimenti

2.1 Contenuto e sviluppo di GPD

Il Gourmet Photography Dataset [19] è stato la base su cui sono stati sviluppati e sperimentati diversi algoritmi per l'estrazione delle feature al fine di studiare l'estetica delle 24000 immagini di cibo che lo compongono. Le fotografie vengono classificate in maniera binaria in base alla loro estetica, quindi le due classi sono rispettivamente:

- **Classe positiva**, la quale contiene 13088 immagini
- **Classe negativa**, la quale contiene 10912 immagini

In Figura 2.1 e Figura 2.2 sono riportati alcuni esempi delle immagini contenute nel GPD per ognuna delle due classi.



Figura 2.1: Esempio di alcune immagini del GPD che a cui è associata la label negativa

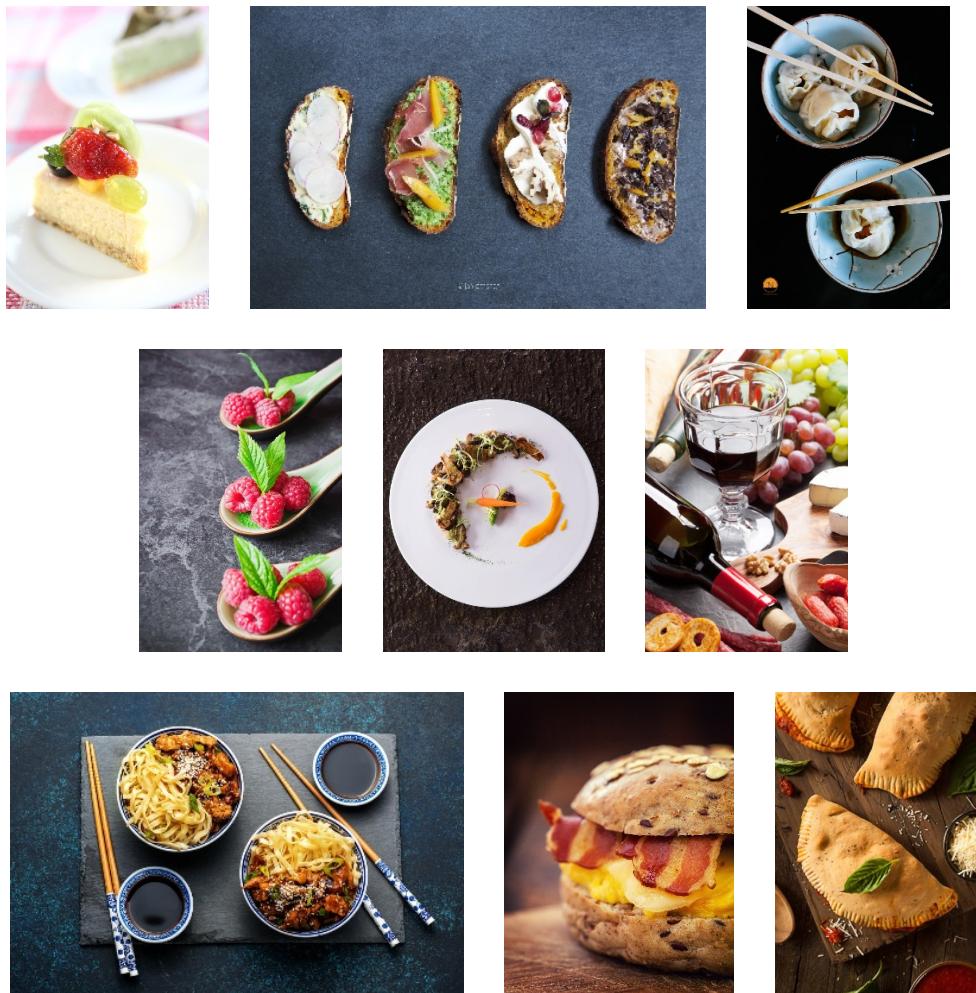


Figura 2.2: Esempio di alcune immagini del GPD che a cui è associata la label positiva

Inoltre questo dataset relativo alla valutazione estetica applicata alle immagini di cibo è stato il primo grande insieme di immagini che unissero la valutazione estetica all'ambito del cibo, in quanto in precedenza erano presenti in letteratura dataset di cibo e delle valutazioni delle prestazioni relative alla classificazione dei vari cibi, ma non c'erano dataset relativi alle due tematiche insieme.

Per creare il GPD sono state necessarie 3 fasi:

- 1. Raccolta delle immagini**, in modo tale che siano il più varie possibile sia nel contenuto che nelle condizioni di luminosità e colori, scartando le immagini duplicate e svolgendo un pre-processing che elimini bordi inutili e ruoti correttamente le immagini.

2. **Scrittura delle label corrispondenti**, considerando il problema dell'estetica come un problema di tipo binario. Si hanno N coppie $\{I_i, \hat{y}_i\}_{i=1}^N$, dove $\hat{y}_i \in \{0,1\}$ è la label associata all'immagine i -esima I_i e N indica la numerosità delle immagini, in questo caso pari a 24000. Le immagini sono state valutate da 57 lavoratori del AMT (Amazon's Mechanical Turk), i quali hanno avuto anche la possibilità di non esprimere un giudizio sulle immagini che secondo loro erano ambigue, poiché è necessario che le label risultanti da questo passaggio siano significative. In particolare quando un'immagine viene saltata tre volte essa viene automaticamente scartata e non viene più riproposta ad altri utenti. Uno schema del procedimento di valutazione appena illustrato è visibile in Figura 2.3.

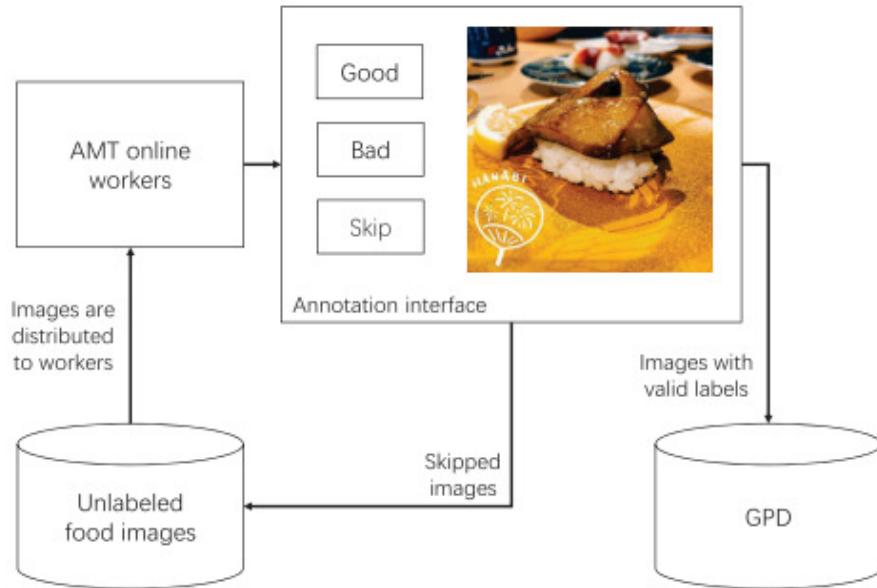


Figura 2.3: Schema tratto dal lavoro originale [19] che illustra la pipeline della valutazione delle immagini del GPD da parte di 57 lavoratori del AMT (Amazon's Mechanical Turk)

3. **Approvazione delle label da parte di otto fotografi esperti**, se almeno quattro di essi sono d'accordo con la label assegnata dagli utenti allora essa viene mantenuta, altrimenti la label è considerata ambigua e l'immagine viene eliminata.

Le immagini verranno poi divise in modo randomico in:

- **Training set**: 21600 immagini, di cui

- 11779 positive
- 9821 negative
- **Test set:** 10912 immagini, di cui
 - 1309 positive
 - 1091 negative

2.2 Test e risultati

Nel lavoro originale [19], sono state provate varie combinazioni di feature e classificatore SVM (Support Vector Machine) con risultati di accuratezza sul test set abbastanza bassi, come è visibile nella prima parte della Tabella 2.1. In particolare le combinazioni sono le seguenti:

- **SVM e colore.** Il colore è una caratteristica molto significativa quando si parla di estetica, in questo caso il colore è stato codificato come un istogramma con 128 bin per i canali RGB. Inoltre i valori sono stati normalizzati rispetto alla media e alla varianza come segue:

$$x' = \frac{x - \mu}{\sigma} \quad (2.1)$$

In particolare μ indica la media, σ indica la varianza e x' indica il valore x a cui è stata applicata la normalizzazione.

- **SVM e feature GIST.** Queste feature sono di tipo globale, in particolare viene estratto un array di feature con lunghezza pari a 512 partendo da una immagine a livelli di grigio di dimensione pari a 256x256. Anche in questo caso i valori sono stati normalizzati rispetto alla media e alla varianza come indicato al punto precedente in Formula (2.1).
- **SVM e feature VGG.** Viene estratto un array di feature con lunghezza pari a 4096 dal penultimo layer di una rete VGG-16. In questo caso sono stati usati tre diversi modelli con contenuto semantico differente: VGG-objects, VGG-scenes e VGG-foods.

L'accuratezza cresce utilizzando delle reti neurali convoluzionali supervisionate, inizializzate con il dataset ImageNet [5] e utilizzando la cross-entropy per minimizzare la perdita e ottimizzare il modello. Nel training vengono utilizzate immagini scalate rispetto al lato più corto, tagliate e specchiate orizzontalmente in modo casuale al fine di aumentare il dataset disponibile.

Solution	Training Set	Test Set
SVM classifier		
SVM + color	72.4	63.3
SVM + GIST	78.1	64.4
SVM + VGG-object	90.8	74.7
SVM + VGG-scenes	86.8	72.4
SVM + VGG-foods	90.4	74.1
Vanilla CNNs		
AlexNet	89.1	88.6
VGG-16	90.6	87.2
InceptionV2	94.0	90.1
ResNet-18	93.3	89.7
CNNs for aesthetic assessment		
MP _{ada} [18]	94.6	90.4
ResNet-18 with regularization		
ResNet-18 + aug	93.6	89.9
ResNet-18 + LSR[23]	95.6	90.2
ResNet-18 + σ_T [6]	94.1	89.4
ResNet-18 + ASR[19]	95.0	90.7

Tabella 2.1: Livelli di accuratezza e risultati raggiunti nel lavoro originale [19] sul training set e sul test set utilizzando diversi approcci: combinazioni del classificatore SVM con diverse feature, reti neurali e reti neurali combinate con metodi di regolarizzazione

Dagli esperimenti è emerso che la dimensione del GPD è sufficiente per classificare in maniera binaria l'estetica dei cibi, senza ricorrere a tecniche complesse di aumento delle immagini del dataset, inoltre il metodo di regolarizzazione ASR (Adaptive Smoothing Regularization) appositamente sviluppato [19] risulta il migliore per quanto riguarda la confidenzialità dei risultati.

Le migliori performance, come è visibile in Tabella 2.1, sono ottenute tramite l'utilizzo del GPD insieme alle reti neurali, in particolare con la ResNet-18 combinata con ASR. Attraverso altri metodi di regolarizzazione è possibile ottenere risultati di accuratezza sul test set attorno al 90%, anche se con ASR è possibile ottenere una maggiore flessibilità.

Un altro risultato degno di nota è legato alla semantica degli oggetti, in questo caso del cibo, quando bisogna valutarne l'estetica in quanto è ben visibile in Tabella 2.1 che utilizzando il classificatore SVM combinato con le feature appartenenti a VGG-scenes l'accuratezza, sia sul test set che sul training set, è minore rispetto a

quando si utilizza VGG-object o VGG-foods e ciò evidenzia che la semantica degli oggetti è significativa quando si vuole condurre una analisi sull'estetica di essi.

Nel lavoro originale [19], a seguito di un ulteriore esperimento con 825 fotografie nuove sono state tratte alcune conclusioni, in particolare 50 candidati selezionati per valutare le immagini hanno ottenuto risultati coerenti rispetto a quelli ottenuti a livello teorico e ciò è particolarmente significativo, in quanto un buon modello per la valutazione estetica deve avere buone capacità di generalizzazione. Inoltre è emerso che le reti neurali supervisionate e parzialmente riaddestrate con il GPD posseggono questa grande capacità di generalizzazione, ciò porta a dimostrare la grandezza e l'importanza dei risultati ottenuti con l'uso del Gourmet Photography Dataset.

Un'ulteriore evidenza significativa è che la valutazione estetica delle fotografie etichettate come negative è stata, in accordo con i risultati ottenuti, molto più semplice rispetto a quella delle immagini etichettate come positive, poiché gli utenti erano maggiormente in accordo tra loro e questo è il motivo che sta alla base del fatto che il GPD è sbilanciato con un maggior numero di immagini positive rispetto a quelle negative.

3

Dataset proposto

3.1 Motivazione e composizione

Per creare questo nuovo dataset sono state raccolte 146 nuove immagini, di cui 85 scattate da un fotografo professionista e 61 scattate da utenti non professionisti. È stato poi estrapolato da esse un sottoinsieme di 100 immagini, rispettivamente 50 professionali e 50 amatoriali, per poterle usare per valutare la capacità di generalizzazione della rete neurale che verrà usata successivamente.

Il motivo per cui si è scelto di proporre un nuovo dataset è stato quello di conoscere le prestazioni ottenute con la rete neurale, precedentemente utilizzata con il test set estratto dal GPD, e un nuovo set di immagini che la rete non avesse mai visto. Una particolarità evidente di questo dataset è il fatto che sia composto sia da fotografie professionali che da fotografie amatoriali, questo perché si è scelto di investigare il tema dell'estetica cercando di capire se la modalità di acquisizione e la tecnica fotografica possano influenzarne la percezione.

Le fotografie professionali sono state scattate nel corso degli anni da un fotografo professionista con macchine fotografiche Canon e principalmente con una lente a lunghezza focale fissa (50mm, 85mm o 100mm), che in questo caso permette di mettere a fuoco il soggetto e ottenere un effetto sfocato nel background, inoltre sono state precedentemente selezionate e contengono enhancement in quanto sono state scattate per menù e social network di ristoranti ed eventi perciò sono state necessarie queste operazioni prima di consegnarle agli utilizzatori finali.

Le fotografie amatoriali provengono da viaggi ed esperienze personali, sono

state scattate principalmente da me in prima persona con iPhone 8 e iPhone 12 Pro oppure, in minima parte, da amici e non contengono enhancement di alcun tipo.

3.2 Assegnamento delle groundtruth

Inizialmente si pensava che le fotografie professionali fossero esteticamente belle, mentre quelle scattate dagli utenti non lo fossero, ma per confermare questa ipotesi è stato richiesto l'intervento di 41 utenti che valutassero le immagini in modo tale da ottenere le vere e proprie groundtruth, le quali potevano coincidere o meno con quelle ipotizzate inizialmente.

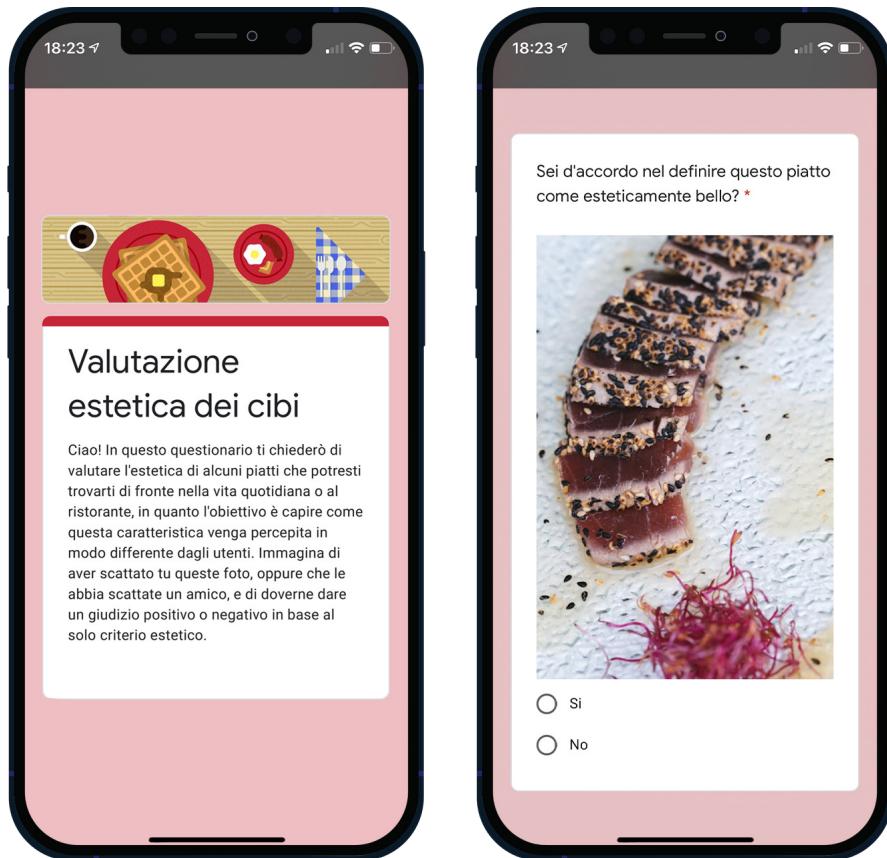
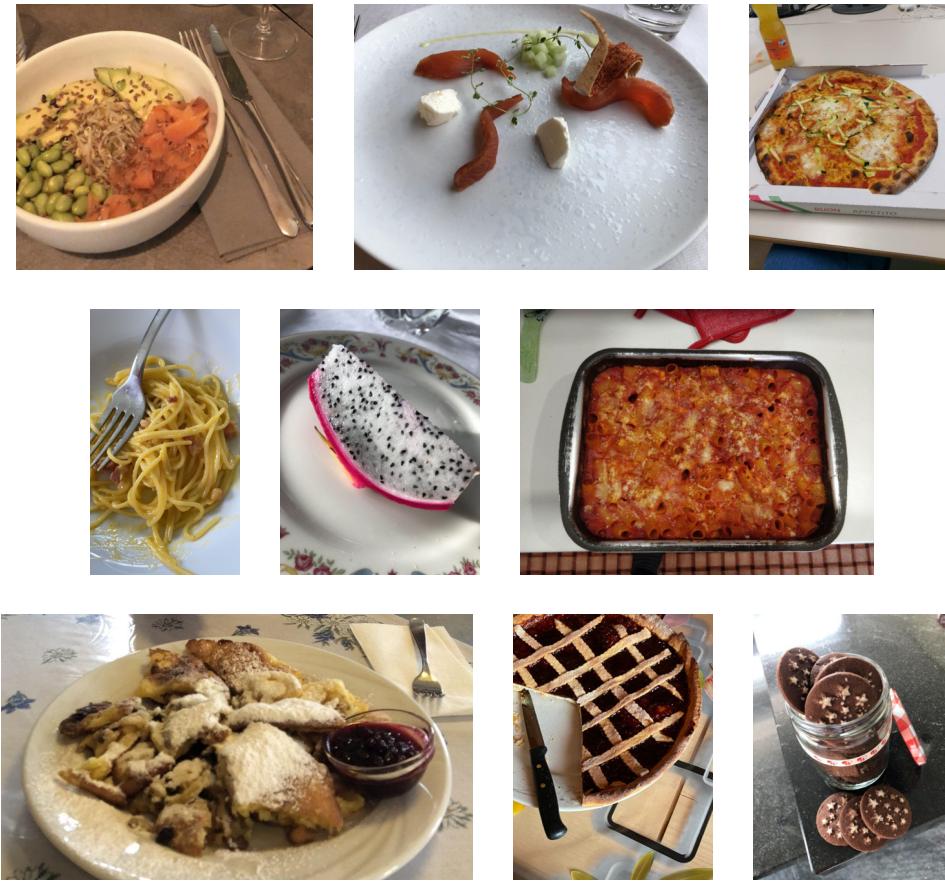


Figura 3.1: Esempio di una delle domande del questionario che è servito per far valutare a 41 utenti le immagini del dataset proposto e ottenere le label di ground-truth

Prendendo ispirazione dallo stato dell'arte e da altri lavori sull'estetica [19], è stato creato un questionario all'interno del quale gli utenti hanno votato se, in

base alla loro opinione e alla loro percezione dell'estetica, i piatti potevano essere considerati esteticamente belli o meno e, quindi, appartenere alla classe positiva o negativa. In base a questi voti sono state assegnate le label, in particolare per ogni immagine I_i la corrispondente label binaria \hat{y}_i è data dalla maggioranza dei voti relativi a quella specifica immagine. Un esempio di una domanda del questionario è visibile in Figura 3.1.

Di seguito, in Figura 3.2, sono riportate tutte le 27 immagini del dataset che sono risultate avere una groundtruth negativa, mentre in Figura 3.3 sono riportate tutte le 73 immagini che hanno groundtruth positiva. Si noti che nonostante si sia partiti da 50 immagini professionali e 50 amatoriali alla fine la numerosità di immagini con groundtruth negativa è minore di quella con groundtruth positiva, in quanto gli utenti hanno valutato positivamente anche alcune fotografie non professionali, ipoteticamente non concentrandosi sulla fotografia in sé bensì sul cibo stesso e pensando se quello specifico cibo fosse di loro gradimento. In minima parte ci sono stati anche casi in cui fotografie professionali sono state valutate negativamente.



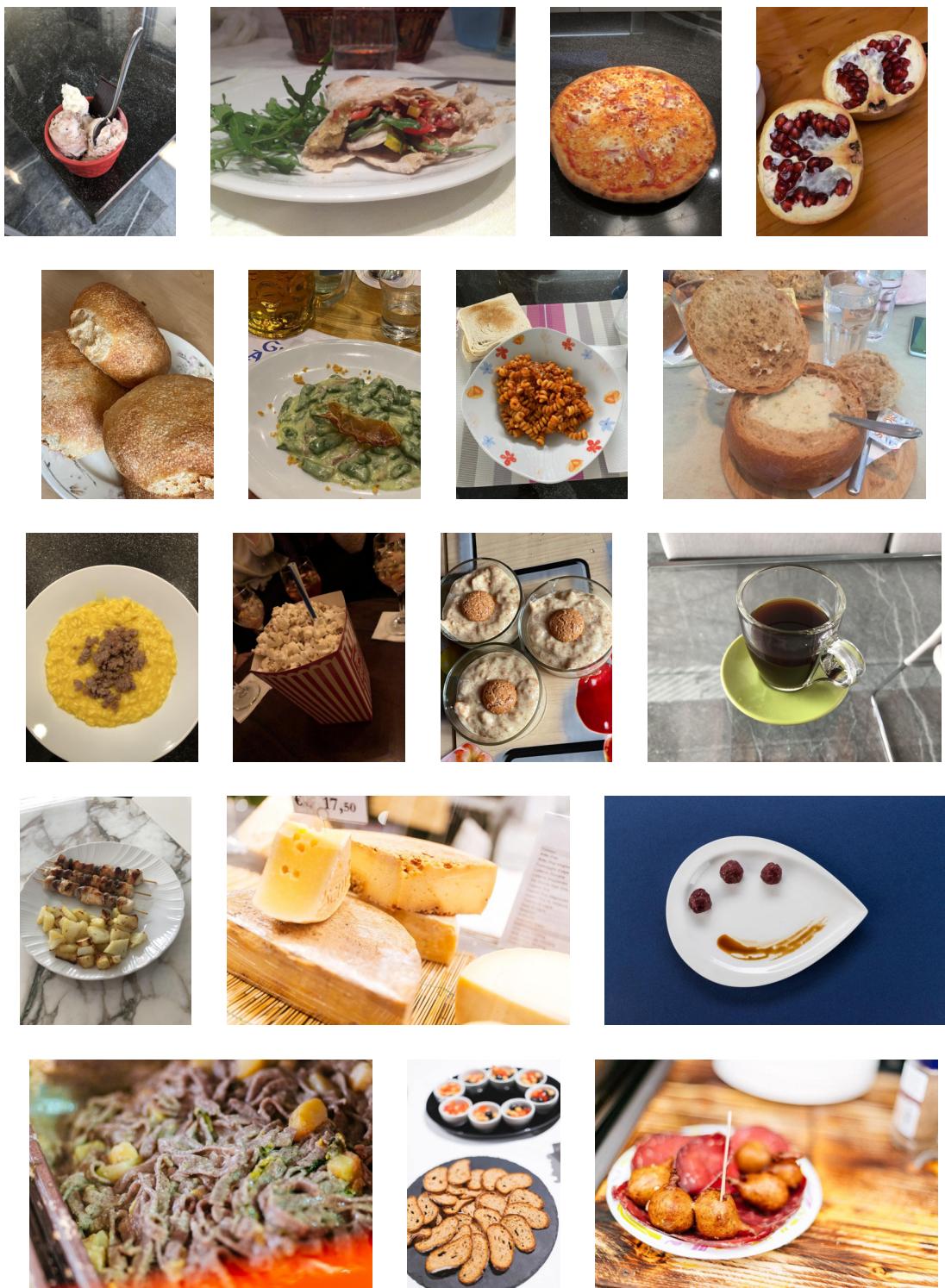
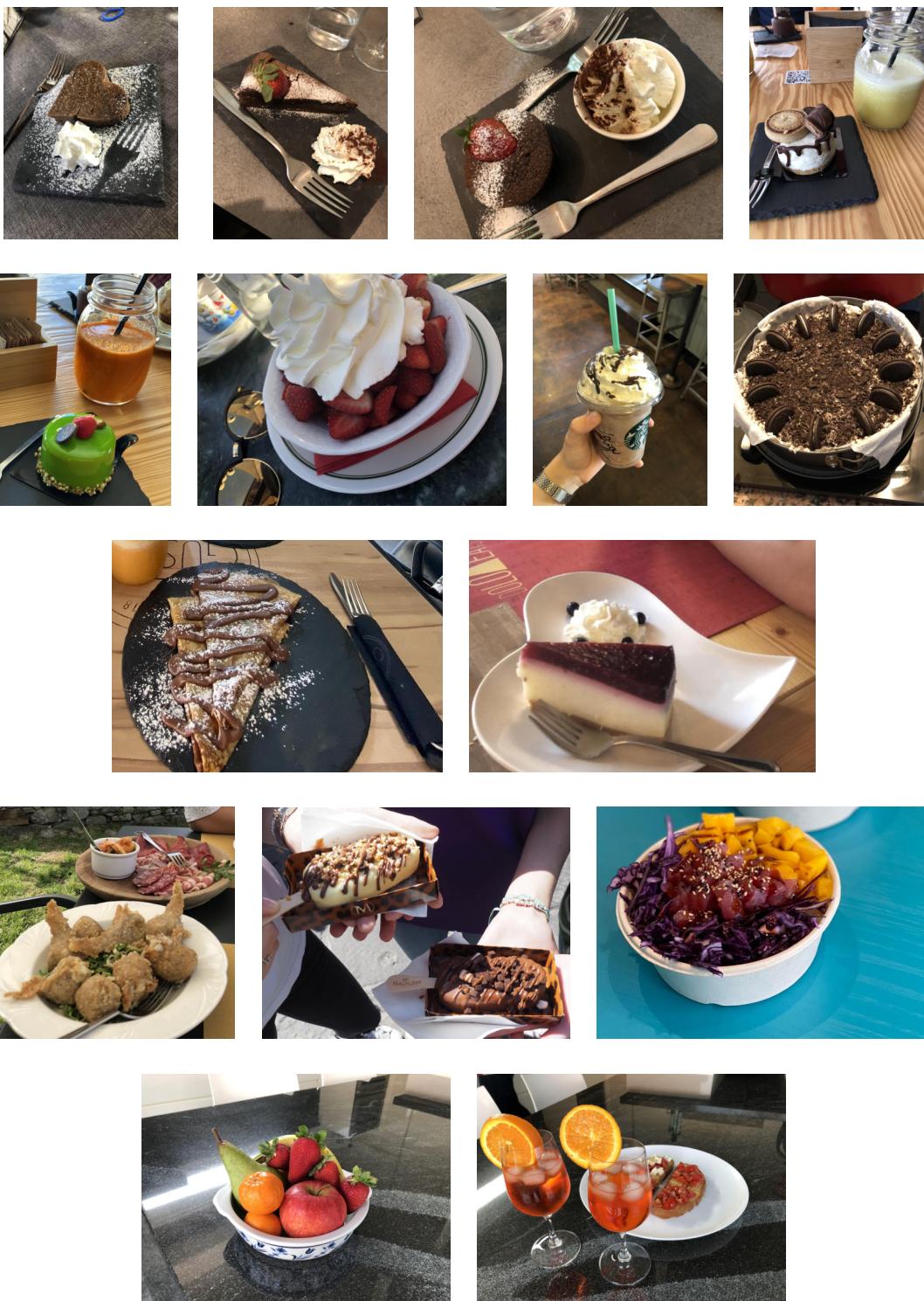
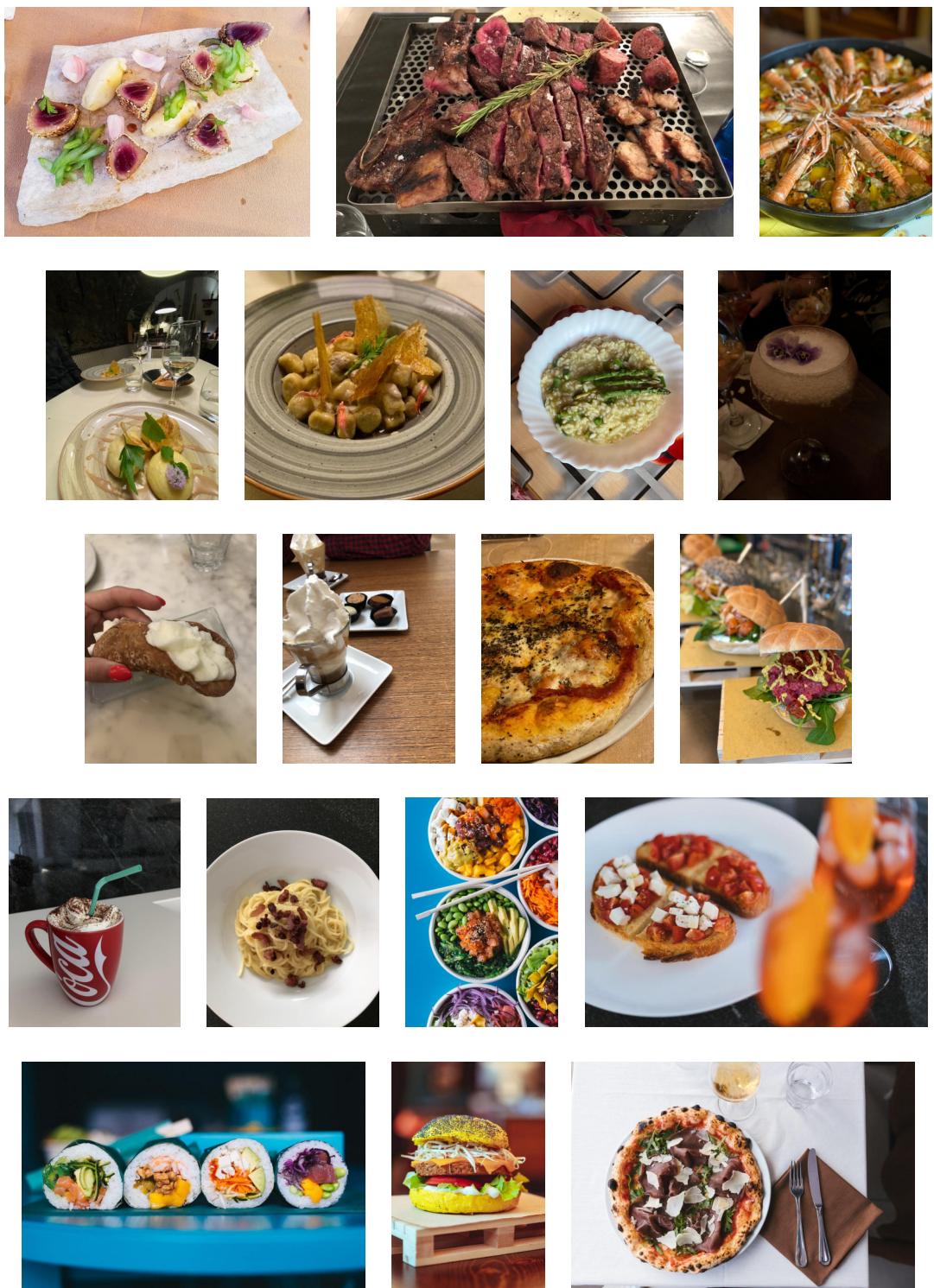
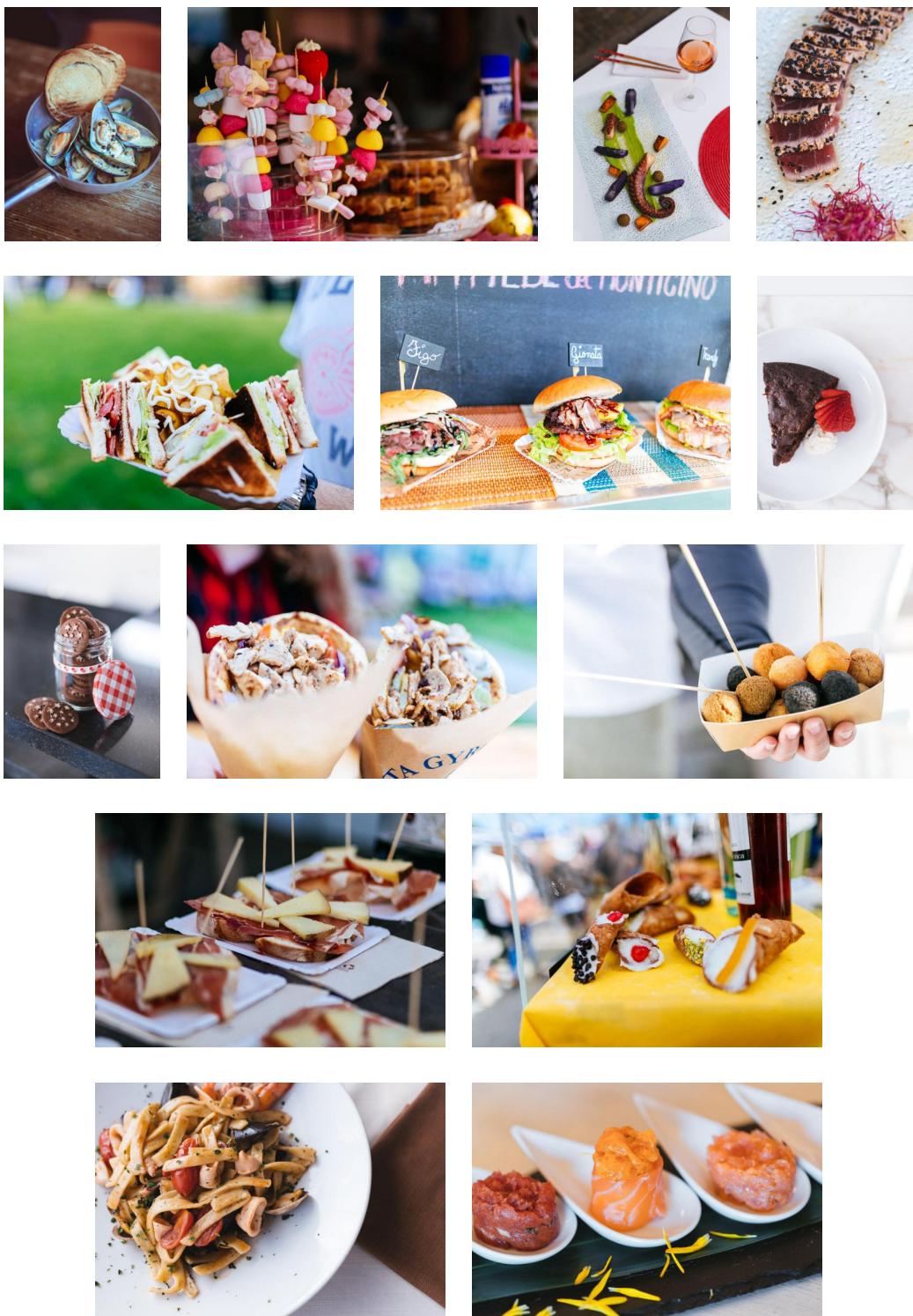
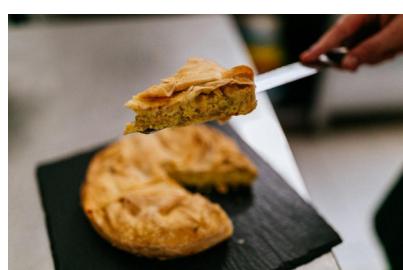
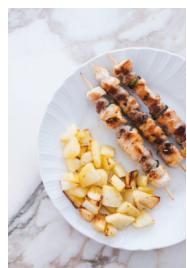


Figura 3.2: Immagini del dataset a cui è stata assegnata una groundtruth negativa tramite i voti degli utenti









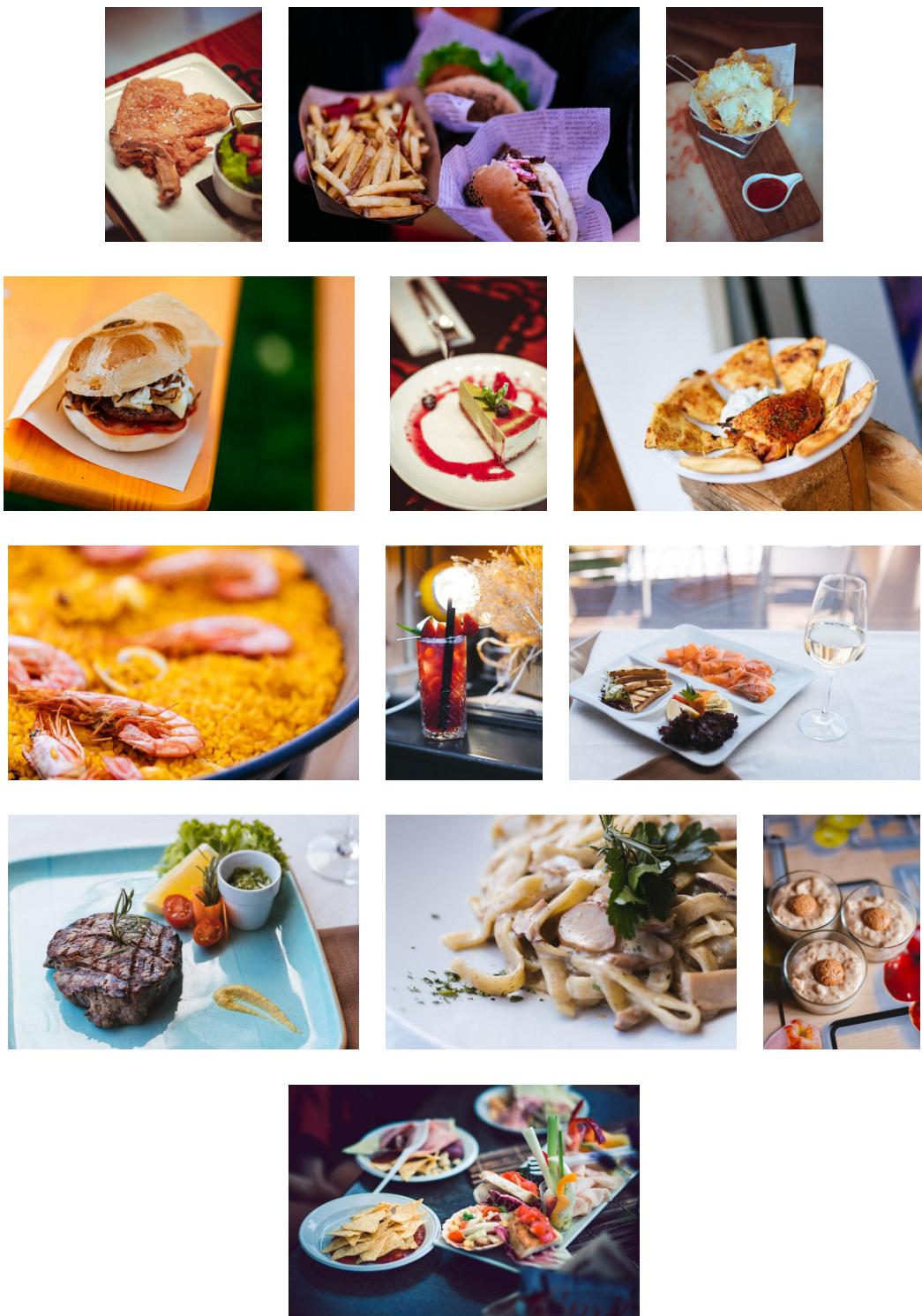


Figura 3.3: Immagini del dataset a cui è stata assegnata una groundtruth positiva tramite i voti degli utenti

Al fine di identificare cosa avesse portato gli utenti a fornire una determinata valutazione è stato creato un secondo questionario analogo al precedente, con la differenza che veniva richiesto loro di motivare il perché di ogni risposta fornita. Un esempio di domanda è riportato in Figura 3.4.

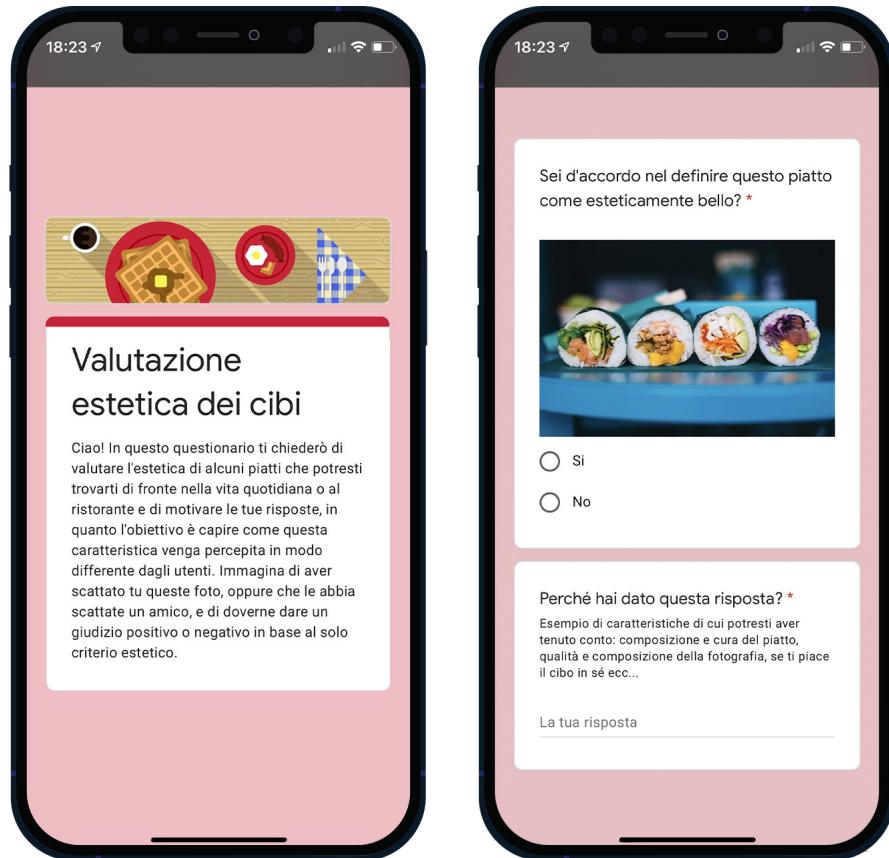


Figura 3.4: Esempio di una delle domande del questionario che è servito per far valutare a 11 utenti le immagini del dataset proposto e a motivare il perché di tale valutazione al fine di comprendere su che aspetti dell'immagine si fossero focalizzati

Il questionario è stato completato da un sottoinsieme di 11 utenti, dai quali è emerso che un cibo esteticamente bello deve essere ben disposto all'interno del piatto, deve essere appetitoso agli occhi di chi lo osserva e se i colori catturano l'attenzione è più probabile che venga valutato positivamente. Inoltre è stata confermata l'ipotesi precedente, ovvero che alcuni utenti in diversi casi hanno valutato positivamente una fotografia non concentrandosi sulla tecnica fotografica utilizzata, ma sul gusto del cibo stesso o sulla sua presentazione.

Ad esempio un utente, per suo gusto personale, non ha apprezzato una fotografia di carne grigliata poiché ipoteticamente potrebbe essere vegetariano oppure

semplicemente potrebbe non piacergli questo cibo in particolare.

In Figura 3.5 viene mostrata una immagine che ha ottenuto 11 voti negativi su 11, le principali motivazioni sono la scarsa accuratezza nella preparazione e composizione del piatto, il fatto che il cibo è già stato in parte mangiato e che si nota la scarsa qualità della fotografia in quanto essa è leggermente sfocata.



Figura 3.5: Esempio di una immagine che ha ottenuto solo voti negativi nel secondo questionario, principalmente a causa del fatto che il cibo non ha una composizione accurata, è già stato mangiato e la fotografia è leggermente sfocata

In Figura 3.6 viene mostrata una immagine che ha ottenuto 11 voti positivi su 11, essa risulta molto colorata e il contrasto cromatico tra i colori dei poke e il tavolo attira l'attenzione di chi osserva, inoltre la composizione della fotografia e del cibo stesso è molto curata e attenta.



Figura 3.6: Esempio di una immagine che ha ottenuto solo voti positivi nel secondo questionario, gli utenti si sono concentrati principalmente sui colori molto vivaci che formano un bel contrasto cromatico con il tavolo e sulla composizione accurata dei poke

4

Metodologie per la classificazione

4.1 Feature hand-crafted

Dopo aver suddiviso le immagini del dataset GPD come già illustrato nel Capitolo 2 sono stati osservati i vari tipi di feature in letteratura, al fine di capire quali utilizzare. Le feature possono essere suddivise in 4 principali categorie [20]:

1. **Feature relative al colore**, ad esempio combinazioni lineari dei tre canali degli spazi colore HSV o HSL oppure la media di essi.
2. **Feature relative alla texture**, ad esempio lo studio degli edge, LBP (Local Binary Pattern), CEDD (Color and Edge Directivity Descriptor) o QHist ovvero l'istogramma dei canali RGB quantizzati a 16 livelli ciascuno.
3. **Feature relative alla composizione**, ad esempio lo studio del blur, delle linee tramite la trasformata di Hough, l'aspect ratio o la regola dei terzi.
4. **Feature relative al contenuto**, ovvero la ricerca di determinati oggetti o aree in una immagine, ad esempio un volto oppure la pelle.

4.2 Un'ulteriore suddivisione delle immagini

Dopo aver analizzato più a fondo i datastore, ovvero le strutture utilizzate in MATLAB [12] per gestire i vari set di immagini, è stato scelto di lavorare con set bilanciati in modo da avere una probabilità equa per le due classi.

In particolare le immagini sono state ulteriormente divise come segue:

- **Set di Rigetto:** 2176 immagini positive che portavano le due classi ad essere sbilanciate, in particolare questo set di immagini non verrà più utilizzato nel proseguo del lavoro. Si avranno 21824 immagini in totale, ovvero la differenza tra il numero totale di immagini (24000) e il numero di immagini appartenenti a questo set.
- **Training set:** 17678 immagini, ovvero il 90% del 90% del nuovo numero totale di immagini (21824). Si prende una prima volta il 90% delle immagini poiché il restante 10% sarà poi parte del Test set, successivamente se ne prende ancora il 90% perché il restante 10% sarà assegnato al Validation set. In particolare le immagini del Training set saranno divise come segue:
 - 8839 positive
 - 8839 negative
- **Test set:** 2182 immagini, ovvero il restante 10% dopo la prima divisione, di cui
 - 1091 positive
 - 1091 negative
- **Validation set:** 1964 immagini, ovvero il restante 10% dopo la seconda divisione del Training set, di cui
 - 982 positive
 - 982 negative

4.3 Feature estratte da una rete neurale

Per cercare di migliorare i risultati ottenuti con le feature hand-crafted si è scelto di utilizzare una rete neurale da cui estrarre delle feature di più alto livello e passarle a un classificatore, in questo caso un SVM. La rete scelta per questa fase è stata la ResNet-18 poiché era quella più utilizzata in altri esperimenti [19] presi come riferimento e che allo stesso tempo otteneva anche buoni risultati.

È stato ipotizzato anche l'uso della VGG-16 ma è stata preferita la ResNet-18 in quanto dopo alcuni test si è notato che era più lenta nel riaddestramento e anche meno utilizzata negli esperimenti presi come riferimento.

Quando bisogna estrarre le feature da una rete neurale è necessario specificare il layer della rete da cui estrarre, in questo caso con la rete ResNet-18 sono stati

fatti due test, uno con il layer pool5 e uno con il layer res3b_relu. Essi si trovano in parti diverse della rete, infatti il layer pool5 si trova alla fine della rete, mentre il layer res3b_relu si trova più o meno a metà di essa.

4.4 Uso di una rete neurale per l'intera classificazione

4.4.1 Prima implementazione

Al fine di incrementare ulteriormente l'accuratezza ottenuta nelle fasi precedenti del lavoro si è scelto di utilizzare una ResNet-18 pre-addestrata su cui svolgere una operazione di Fine Tuning, ovvero una parziale modifica della rete al fine di adattarla per svolgere un nuovo task. In particolare in questa fase sono stati modificati il Fully Connected layer e il Classification Output layer in modo tale che lavorino con due classi e restituiscano in output solo una di esse.

Nella fase di riaddestramento della rete è stato necessario scegliere se compiere delle operazioni di aumento del dataset disponibile in particolare, traendo ispirazione dal lavoro originale [19], si è scelto di compiere scaling sia lungo l'asse x che lungo l'asse y, traslazione e riflessione rispetto all'asse x in maniera casuale al fine di aumentare il numero di immagini disponibili.

È stato necessario anche impostare alcuni parametri tra cui il numero massimo di epoch, ovvero il massimo numero di cicli di training, la dimensione del batch, ovvero un parametro che definisce il numero di campioni con cui lavorare prima di aggiornare i parametri interni al modello, e il tasso di apprendimento iniziale della rete neurale.

4.4.2 Early Stopping

Durante la fase di riaddestramento della rete si è osservato che, dopo un determinato periodo di tempo, il grafico dell'accuratezza tendeva a rimanere costante e ciò significa che la rete non stava più apprendendo nulla dai dati forniti ad essa. Per questo motivo si è scelto di implementare l'Early Stopping o stop anticipato, il quale permette di fermare il processo di training se l'accuratezza non migliora oppure se la funzione Loss peggiora per un certo numero di epoch consecutive.

In questo caso si è scelto di implementare lo stop basandosi sull'accuratezza, ad esempio se si sceglie di fermare l'addestramento dopo due epoch dove l'accuratezza non si incrementa di fatto si otterranno almeno tre epoch di training, in quanto la rete controllerà se l'accuratezza si è mantenuta minore o uguale a quella della prima epoca nelle due epoch successive e, in caso affermativo, fermerà il processo.

Lo stop anticipato può permettere un risparmio di tempo e risorse durante il Fine Tuning, ma ciò dipende da come si impostano i parametri iniziali e dal numero di epoche dopo le quali si sceglie di fermare il processo.

È importante segnalare che il processo di Early Stopping potrebbe non essere sempre vantaggioso in quanto è possibile ottenere anche casistiche nelle quali, nonostante lo si utilizzi, vengano comunque compiute tutte le epoche di training perché l'accuratezza risulterà sempre in crescita e quindi la rete non fermerà il training in anticipo, poiché si cerca sempre di ottenere il miglior risultato possibile.

4.5 Valutazione del dataset proposto

Per tutto il resto del lavoro è stata utilizzata la rete ResNet-18 riaddestrata con Early Stopping che ha ottenuto migliori risultati sia sul test set che sul validation set, ovvero quella con Initial Learn Rate pari a 0.0005, stop anticipato dopo due epoche delle sei possibili e dimensione del batch pari a dieci.

Con questa rete sono state valutate le immagini del dataset proposto, prestando attenzione al fatto che in MATLAB alcune immagini erano visualizzate ruotate di 90 gradi a causa di metadati contenuti all'interno delle immagini stesse, per cui è stato necessario un processing mirato su alcune immagini al fine di avere tutto il dataset visualizzato correttamente.

Inizialmente, come già specificato, si era pensato di utilizzare come ground-truth delle label ricavate secondo il criterio che una fotografia professionale fosse esteticamente bella, e quindi avesse label positiva, mentre una fotografia scattata da un utente non lo fosse, e quindi avesse label negativa. Ciò è parso poco sensato, in quanto osservando le immagini del dataset proposto riportate nel Capitolo 3 in Figura 3.2 e Figura 3.3 si può notare che alcune immagini non professionali sono state comunque etichettate come positive dagli utenti nel questionario a loro sottoposto e viceversa, ciò sta a significare che non sempre c'è una corrispondenza tra tecnica nello scatto ed estetica di esso, anche se generalmente molte fotografie professionali sono percepite come esteticamente belle poiché utilizzano colori, contrasti e tecniche che gli utenti apprezzano molto.

A seguito di questa considerazione si è quindi scelto di utilizzare come ground-truth per la valutazione le label risultanti dal questionario, in quanto esse catturano in maniera particolare la percezione dell'estetica del dataset proposto da parte di un particolare gruppo di 41 utenti che sono stati presi come campione per questo esperimento. In particolare le immagini saranno distribuite nelle due classi come segue:

- **Groundtruth appartenente alla classe positiva:** 73 immagini
- **Groundtruth appartenente alla classe negativa:** 27 immagini

4.6 Grad-CAM per capire le predizioni

Al fine di comprendere quali porzioni di un'immagine vengano maggiormente prese in considerazione dalla rete per determinare la label predetta è stata utilizzata la tecnica del Gradient-weighted Class Activation Mapping (Grad-CAM) [17], la quale evidenzia in rosso le aree più significative e in blu quelle meno significative per la predizione della rete fornendo una visualizzazione facilmente interpretabile da chi la osserva ed essendo applicabile a molte famiglie di reti neurali convoluzionali.

In particolare la tecnica in questione utilizza il gradiente relativo al concetto che si è scelto come target nell'immagine, il quale entra negli ultimi layer convoluzionali, con l'obiettivo di produrre una mappa che mostri le aree più significative e quelle meno significative relative al tema scelto. Si utilizzano proprio questi layer poiché le feature convoluzionali conservano informazioni spaziali, le quali vengono perse all'interno del Fully Connected layer e quindi gli ultimi layer convoluzionali prima di esso permettono di conciliare una buona semantica con queste informazioni che sono molto significative se, come in questo caso, si vuole analizzare quale area dell'immagine è più rilevante per la classificazione di essa da parte di una rete neurale.

Un esempio di applicazione della tecnica Grad-CAM ad una immagine è riportato in Figura 4.1.

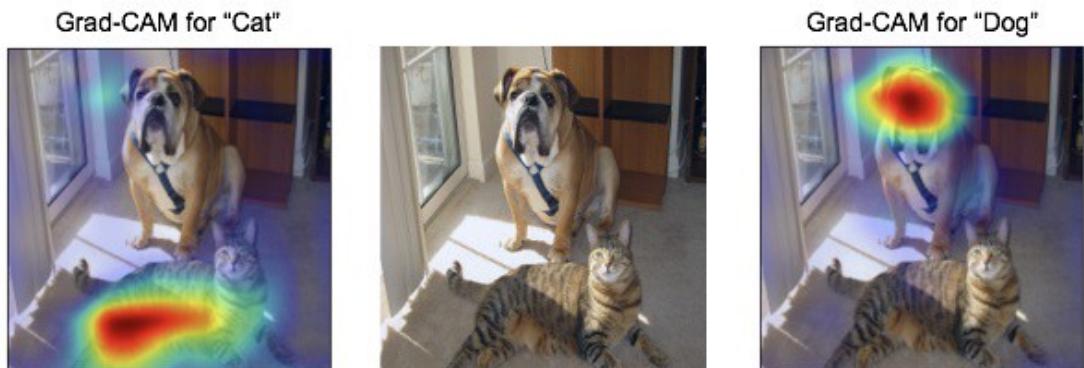


Figura 4.1: Esempio di una immagine a cui è stata applicata la tecnica Grad-CAM focalizzandosi su due differenti classi di oggetti, in questo caso prima sulla classe "Gatto" e poi sulla classe "Cane", il quale è stato tratto dal lavoro originale sulla tecnica Grad-CAM [17]

Osservando le immagini ricavate con la tecnica appena illustrata applicata alle immagini del dataset proposto sembrava potesse esserci una correlazione tra la predizione della rete e la porzione di immagine che veniva considerata più significativa, ovvero quella colorata di rosso.

Al fine di avere una misura oggettiva che potesse aiutare a comprendere se ci fosse o meno questa correlazione è stato necessario compiere alcuni passaggi preliminari:

- **Creazione di maschere binarie**, le quali avessero in bianco il cibo e in nero il background, un esempio è riportato in Figura 4.2. Esse sono state realizzate con l'ausilio di Photoshop, ma sarebbe andato bene anche un altro editor di immagini.

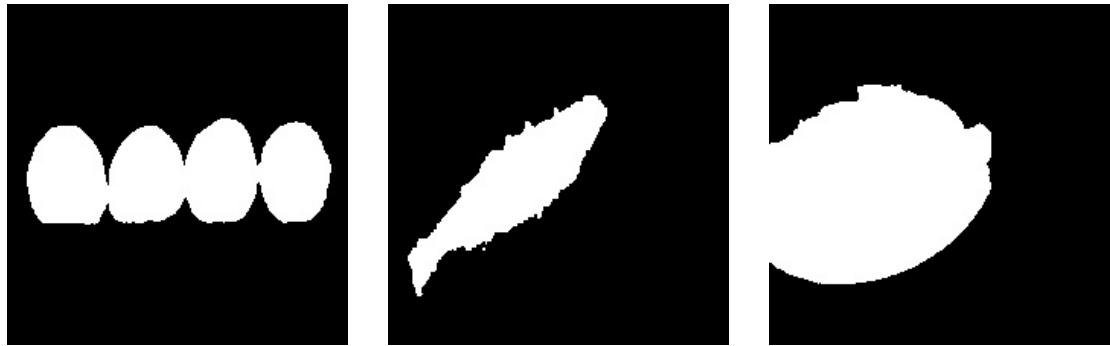


Figura 4.2: Esempi di maschere binarie dei cibi, le quali mostrano in bianco il cibo e in nero il background

- **Calcolo del rapporto tra area della maschera e area totale dell'immagine**, tale rapporto è stato espresso in percentuale ed è indicato con A_m .
- **Calcolo del rapporto tra energia che ricade nella maschera ed energia totale**, dove con energia si intende l'intensità della mappa Grad-CAM per ogni pixel. Questo passaggio è stato fatto ponendo attenzione a normalizzare i valori delle mappe Grad-CAM rispetto al valore minimo e massimo, in modo tale da ottenere valori tra zero e uno, come segue:

$$x' = \frac{x - \min}{\max - \min} \quad (4.1)$$

In particolare anche questo rapporto è stato espresso in percentuale per comodità e viene indicato con E_m .

- **Calcolo dell'indicatore di concentrazione dell'energia**, ovvero calcolo del seguente rapporto:

$$C = \frac{E_m}{A_m} \quad (4.2)$$

Tale valore per definizione sarà un valore decimale e non una percentuale, poiché è definito come rapporto tra due percentuali. Si è scelto di mantenerlo in questa forma per comodità nel visualizzare grafici che lo rappresentassero al fine di correlarlo alle predizioni della rete.

5

Risultati ottenuti

5.1 Risultati ottenuti sul dataset GPD

5.1.1 Feature hand-crafted

Per prima cosa sono stati analizzati i risultati ottenuti con le feature hand-crafted, ovvero le più semplici, i quali sono riportati in Tabella 5.1.

Feature	Classificatore	Training Set	Test Set
LBP	SVM	0.7276	0.6987
CEDD	SVM	0.6788	0.6529
QHist	SVM	0.6931	0.6904
LBP, CEDD, QHist	SVM	0.6766	0.6667
Media H, Media S, Intensità	SVM	0.5813	0.6075
Pleasure, Arousal, Dominance	SVM	0.5688	0.5850
Media H, Media S, LBP	SVM	0.7357	0.7446
Media H, Media S, CEDD	SVM	0.6772	0.6854
Pleasure, Arousal, Dominance, LBP	SVM	0.7277	0.7383
Pleasure, Arousal, Dominance, CEDD	SVM	0.6821	0.6867

Tabella 5.1: Livelli di accuratezza ottenuti sul training set e sul test set utilizzando diverse combinazioni di feature e il classificatore SVM

In questo caso sono state usate principalmente feature legate al colore e alla texture e il miglior risultato sul test set è stato ottenuto combinando la media di canali H e S insieme a LBP, quindi unendo descrittori legati al colore e alla texture. Inoltre in tabella sono stati riportati i risultati su test set e su training set, i quali sono stati divisi come illustrato nel Capitolo 2, per poter fare un immediato confronto con i risultati ottenuti nel lavoro originale [19], i quali sono stati riportati in Tabella 2.1.

5.1.2 Feature estratte da una rete neurale

Di seguito, in Tabella 5.2, sono riportati i risultati ottenuti con feature estratte da due layer differenti di una rete ResNet-18 e, in particolare, non c'è una grande differenza di accuratezza estraendo le feature da un layer che si trova più in profondità all'interno della rete come il layer pool5 rispetto a uno che si trova meno in profondità come il layer res3b_relu, anzi l'accuratezza è leggermente più elevata nel secondo caso.

Di seguito, così come negli studi successivi, verranno riportati i livelli di accuratezza ottenuti sul test set e sul validation set come spiegato precedentemente in Sezione 4.2 del Capitolo 4 al fine di utilizzare dei dataset che avessero una numerosità equa di immagini in ognuna delle due classi.

Rete	Classificatore	Layer	Test Set	Validation Set
ResNet-18	SVM	pool5	0.8350	0.8452
ResNet-18	SVM	res3b_relu	0.8391	0.8478

Tabella 5.2: Livelli di accuratezza ottenuti sul validation set e sul test set utilizzando il classificatore SVM con delle feature estratte da due diversi layer della rete ResNet-18. Nel primo caso le feature sono state estratte dal layer pool5, il quale si trova alla fine della rete, mentre nel secondo caso dal layer res3b_relu, il quale si trova a metà della rete

5.1.3 Uso delle reti neurali per l'intera classificazione

Quando ci si affida completamente all'uso di una rete neurale è necessario differenziare quando si utilizzi la tecnica dell'Early Stopping rispetto a quando non la si utilizzi, ciò porta nella maggior parte dei casi a uno spreco di tempo e risorse poiché vengono eseguite tutte le epoche del riaddestramento della rete anche se, come già anticipato nel Capitolo 4, non è detto che l'uso del Fine Tuning combinato con l'Early Stopping diminuisca il numero di epoche effettivamente svolte in quanto ciò dipende da come vengono impostati i vari parametri e da come cresce l'accuratezza durante l'esecuzione.

I risultati ottenuti senza stop anticipato sono riportati in Tabella 5.3, mentre quelli relativi all'esecuzione con Early Stopping sono riportati in Tabella 5.4.

Rete	Test Set	Validation Set	Initial Learn Rate	Epochs	Batch Size
ResNet-18	0.8882	0.89	0.0003	6	10
ResNet-18	0.8996	0.8905	0.0005	6	10
ResNet-18	0.8731	0.8849	0.00003	6	10
ResNet-18	0.8520	0.8493	0.00001	6	10
ResNet-18	0.8744	0.8829	0.0003	8	10

Tabella 5.3: Livelli di accuratezza ottenuti sul validation set e sul test set utilizzando una procedura di Fine Tuning su una rete ResNet-18. Sono stati riportati anche i parametri utilizzati per inizializzare il numero di epoche, la dimensione del batch e il tasso di apprendimento iniziale della rete

Rete	Test Set	Validation Set	Initial Learn Rate	Epochs	Batch Size	Early Stopping
ResNet-18	0.9051	0.9043	0.0005	6	10	2 epoche
ResNet-18	0.9015	0.8946	0.0003	6	10	2 epoche
ResNet-18	0.8579	0.8569	0.00003	6	10	1 epoca
ResNet-18	0.8437	0.8462	0.00001	6	10	1 epoca
ResNet-18	0.8799	0.8997	0.0003	8	10	2 epoche

Tabella 5.4: Livelli di accuratezza ottenuti sul validation set e sul test set utilizzando una procedura di Fine Tuning su una rete ResNet-18 con Early Stopping dopo una o due epoche, a seconda di quale valore portasse a un maggior risparmio di tempo, anche se nel caso di stop anticipato dopo una sola epoca con Initial Learn Rate pari a 0.00001 non c'è stato alcun miglioramento in quanto sono state eseguite comunque tutte le epoche del training. Sono stati riportati anche i parametri utilizzati per inizializzare il numero di epoche, la dimensione del batch e il tasso di apprendimento iniziale della rete

5.2 Risultati ottenuti sul dataset proposto

5.2.1 Analisi delle groundtruth

Osservando la distribuzione delle groundtruth delle immagini del dataset proposto, la quale è riportata in Sezione 4.5 del Capitolo 4, si osserva che gli utenti hanno valutato positivamente 73 immagini del dataset e negativamente le restanti 27. Questo risultato è sbilanciato verso la classe positiva, nonostante inizialmente siano state utilizzate 50 immagini professionali e 50 scattate da utenti comuni non professionisti, ciò mostra come la percezione dell'estetica dei cibi, ma anche in

qualsiasi altro ambito, non è strettamente legata alla tecnica fotografica utilizzata o alla struttura della fotografia bensì entrano in gioco anche altri fattori soggettivi, tra cui il background culturale, le abitudini alimentari e gli specifici gusti del sottoinsieme di utenti che sono stati chiamati a valutare le immagini del dataset proposto.

5.2.2 Uso di una rete neurale adattata

Come già anticipato nel Capitolo 4 per valutare l'estetica nel dataset proposto è stata utilizzata principalmente la rete neurale che ha ottenuto i migliori risultati ovvero, come riportato in Tabella 5.4, quella che ha un Initial Learn Rate pari a 0.0005 e che utilizza l'Early Stopping dopo due epochhe nelle quali l'accuratezza non migliora.

Utilizzando come groundtruth le label ottenute dal questionario che è stato sottoposto a 41 utenti, la cui distribuzione è stata analizzata precedentemente, e la rete precedentemente citata è stata ottenuta una accuratezza del 72%.

In questa fase è stata molto significativa l'analisi degli errori compiuti dalla rete e la successiva analisi delle immagini generate con la tecnica Grad-CAM [17], in particolare si evince che la rete ha classificato 17 immagini come falsi negativi, ovvero è stata predetta la label negativa mentre la groundtruth era positiva, e 11 falsi positivi, ovvero è stata predetta la label positiva mentre la groundtruth era negativa.

Di seguito, in Figura 5.1 e in Figura 5.2, sono riportate alcune delle immagini del dataset proposto a cui è stata applicata la tecnica Grad-CAM e dove la rete non ha compiuto errori di classificazione.



Figura 5.1: Esempi di immagini a cui è stata applicata la tecnica Grad-CAM dove la rete ha predetto correttamente una label negativa

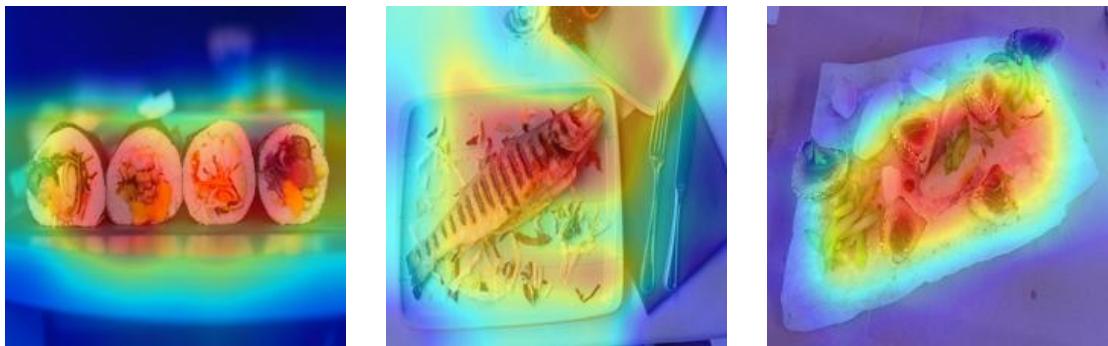


Figura 5.2: Esempi di immagini a cui è stata applicata la tecnica Grad-CAM dove la rete ha predetto correttamente una label positiva

In Figura 5.3 è riportato un esempio di immagine appartenente ai falsi negativi, mentre in Figura 5.4 è invece riportato un esempio di immagine appartenente ai falsi positivi. Entrambe le immagini di esempio sono riportate insieme alle rispettive Grad-CAM al fine di poter mostrare eventuali differenze nella posizione delle aree più significative, in rosso, e di quelle meno significative, in blu, rispetto al cibo. Questa tematica verrà approfondita nella Sezione 5.2.3 insieme ad altre ipotesi scaturite dall’analisi delle Grad-CAM, le quali hanno portato a investigare la correlazione tra l’indicatore di concentrazione dell’energia e le predizioni della rete.



Figura 5.3: Esempio di una immagine per la quale la rete ha predetto la label negativa mentre la groundtruth era positiva, per cui essa rientra tra gli errori e in particolare tra i falsi negativi. A destra viene riportata anche la rispettiva Grad-CAM, dalla quale si nota che le aree che la rete ha considerato più significative per la predizione della label non sono quelle del cibo, bensì appartengono al piatto e allo sfondo

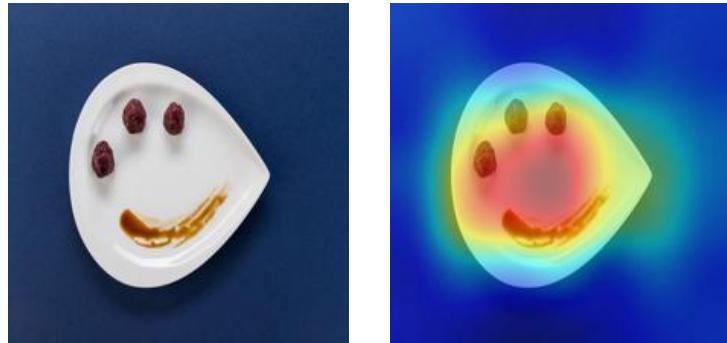


Figura 5.4: Esempio di una immagine per la quale la rete ha predetto la label positiva mentre la groundtruth era negativa, per cui essa rientra tra gli errori e in particolare tra i falsi positivi. A destra viene riportata anche la rispettiva Grad-CAM, dalla quale si nota che le aree che la rete ha considerato più significative per la predizione della label sono effettivamente quelle del cibo e non altre parti dell'immagine

5.2.3 Analisi delle Grad-CAM

Come già evidenziato in precedenza l'analisi delle immagini ricavate con la tecnica Grad-CAM è fondamentale al fine di individuare cosa la rete consideri significativo per l'effettiva predizione della label ed eventualmente per individuare nessi causali tra queste aree, considerate più significative per la predizione, e gli errori compiuti. Infatti osservando le Grad-CAM sembrava che quando l'area più significativa veniva individuata al di fuori del cibo la rete predicesse la label negativa, perciò si è scelto di indagare più a fondo sfruttando l'indicatore di concentrazione dell'energia.

Per confermare quest'ultima ipotesi è stato costruito un grafico, visibile in Figura 5.5, nel quale l'indicatore di concentrazione, riportato sull'asse x e calcolato come espresso in Formula (4.2), viene correlato alle predizioni della rete, riportate sull'asse y. Al fine di rappresentare le label in maniera numerica si è scelto di far corrispondere alla label negativa il valore -1 e alla label positiva il valore 1, in modo tale che la distinzione nel grafico fosse evidente, inoltre si è scelto un grafico di tipo Scatter Plot perché visivamente mostra molto bene la distribuzione dei punti e perché esso dava la possibilità di inserire sugli assi cartesiani due diverse variabili, che in questo caso sono le predizioni e l'indicatore di concentrazione C.

Osservando il grafico è ben visibile che gli elementi della classe positiva sono molto più numerosi di quelli della classe negativa, le immagini sono infatti classificate dalla rete come segue:

- **Classe positiva:** 67 elementi

- **Classe negativa:** 33 elementi

È molto interessante la distribuzione vera e propria di questi elementi, rappresentati nello Scatter Plot come dei pallini, poiché si nota che al di sopra di un certo valore di soglia, che potrebbe essere individuato attorno al valore 1 sull'asse delle x, è possibile separare le due classi commettendo pochi errori. Questa distribuzione mostra una correlazione tra le variabili inserite sugli assi cartesiani così come era stato ipotizzato, al crescere dell'indicatore di concentrazione dell'energia all'interno dell'area corrispondente al cibo cresce la probabilità che venga predetta la label positiva, in particolare man mano che aumenta la concentrazione si osserva che le predizioni diventano tutte appartenenti alla classe positiva.

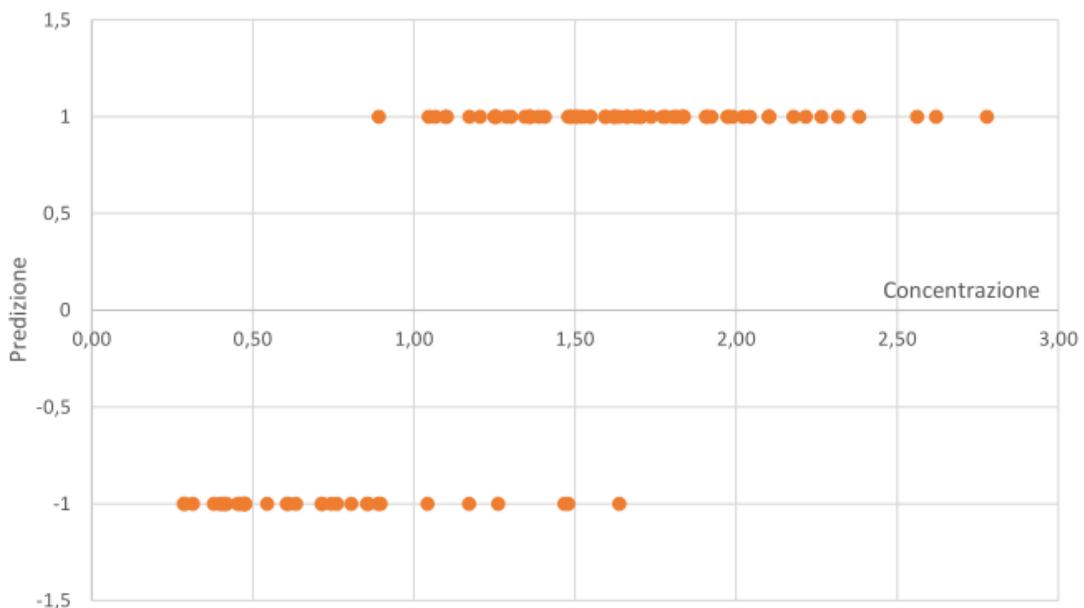


Figura 5.5: Grafico che rappresenta sull'asse delle x l'indicatore di concentrazione per ognuna delle immagini del dataset proposto, mentre sull'asse delle y rappresenta le label predette, in particolare il valore -1 indica la label negativa mentre il valore 1 indica la label positiva

Questo risultato è davvero significativo poiché conferma l'ipotesi fatta in precedenza, mostrando che quando la Grad-CAM si focalizza correttamente sulla porzione di immagine occupata dal cibo si avrà che la concentrazione di energia in quell'area sarà elevata e, quindi, sarà più probabile che la rete predica una label positiva.

Il risultato va però contestualizzato, in determinati casi entrano in gioco fattori di gusto personale degli utenti e fattori legati alla semantica dell'immagine analizzata che possono essere imprevedibili per la rete neurale in quanto, come già sottolineato, l'estetica è una caratteristica puramente soggettiva e che può variare nel tempo, in quanto un utente potrebbe valutare positivamente un cibo e, in un momento successivo, valutare lo stesso cibo in maniera opposta poiché il suo background culturale e il suo bagaglio di esperienze potrebbero avergli fatto modificare la sua personale concezione di bellezza.

6

Conclusioni e sviluppi futuri

La relazione si è basata sul lavoro svolto durante l’esperienza di stage presso l’Ateneo, il cui obiettivo è stato in primo luogo l’analisi di un dataset di immagini di cibo già esistente. Il dataset in questione è il Gourmet Photography Dataset o GPD [19], il quale contiene 24000 immagini di cibi con le relative label, le quali possono appartenere alla classe positiva o alla classe negativa a seconda della valutazione estetica della singola immagine.

Si è scelto di utilizzare diverse feature con grado di astrazione e difficoltà crescente, partendo dalle feature hand-crafted e arrivando fino all’uso di una rete neurale, opportunamente modificata e adattata, per l’intera classificazione delle immagini del dataset. Lo sviluppo di codice per l’intero lavoro è stato in linguaggio MATLAB [12].

Successivamente è stato proposto un dataset creato ad hoc per sperimentare l’utilizzo di una rete neurale appositamente adattata e osservare i risultati ottenuti, confrontandoli con quelli conseguiti con il dataset già esistente. In particolare ci si è poi concentrati sullo studio di ciò che la rete considerava più importante per determinare la label predetta, ovvero una tematica molto interessante e che permette di capire il ragionamento alla base delle valutazioni compiute dalla rete neurale.

A seguito delle analisi svolte durante l’esperienza di stage sul dataset proposto si è osservato che gli utenti tendono a valutare positivamente le immagini, anche se queste fotografie non sono professionali, concentrandosi soprattutto sul cibo e non sulla tecnica fotografica. Inoltre, come già si ipotizzava dopo aver studiato altri lavori [19] già esistenti sul tema dell’estetica dei cibi, si è confermato che l’utilizzo

delle reti neurali ottiene migliori performance rispetto all'utilizzo di descrittori più semplici combinati con un classificatore.

Per quanto riguarda i possibili sviluppi di questa ricerca si potrebbe ipotizzare un sostanziale ampliamento del dataset proposto, portandone la numerosità delle immagini a livelli pari di altri dataset famosi come il GPD, e un ampliamento della popolazione che, tramite il primo dei due questionari, ha fornito le valutazioni delle immagini per determinare le groundtruth utilizzate. Di pari passo con questo ampliamento verrebbe ampliata anche la popolazione che è stata richiamata per completare un secondo questionario, il cui obiettivo era raccogliere le motivazioni degli utenti a seguito della valutazione di ogni singola immagine. L'incremento del numero di utenti coinvolti potrebbe essere particolarmente significativo, in quanto maggiore è il numero di persone che valutano le immagini più sarà possibile osservare il comportamento di esse ed eventualmente modificare la raccolta delle groundtruth o la valutazione delle immagini stesse.

Un possibile sviluppo più a lungo termine potrebbe essere l'integrazione di questo studio in un'applicazione che, data una fotografia di cibo, ne dia una valutazione estetica e fornisca delle motivazioni relative ad essa, in modo tale che ristoranti e chiunque fotografi il cibo possa comprendere se, di fronte a un'immagine valutata negativamente, il problema sia relativo alla fotografia in sé oppure sia legato alla presentazione del cibo. Ciò potrebbe essere utile nell'ottica di creare fotografie non professionali per i social network di un ristorante, un bar oppure semplicemente per chi ama scattare fotografie a ciò che mangia per pubblicarle. Scattare velocemente una foto ed eventualmente applicare qualche filtro di color correction direttamente dallo smartphone, caricarla nell'app e ottenere la valutazione potrebbe essere un processo rapido e semplice prima della pubblicazione di tale fotografia e potrebbe aiutare soprattutto chi usa i social network a livello commerciale per comprendere se l'immagine finale può essere adatta al proprio pubblico su un determinato social network, dato che online gli utenti vengono attirati da fotografie colorate, con contrasti, giochi di luce particolari e, nel caso del cibo, che siano ben strutturate per mostrare qualcosa di appetitoso.

Riferimenti

Bibliografia

- [1] Raffaele Campo, Giuseppe Loporcaro e Fabrizio Baldassarre. «The effects of food aesthetics on consumers. Visual stimuli and food marketing». In: *DIEM: Dubrovnik International Economic Meeting*. Vol. 3. 1. Sveučilište u Dubrovniku. 2017, pp. 553–565 (cit. a p. 9).
- [2] Huiwen Chang et al. «Automatic triage for a photo series». In: *ACM Transactions on Graphics (TOG)* 35.4 (2016), pp. 1–10 (cit. alle pp. 6, 7).
- [3] Kuang-Yu Chang, Kung-Hung Lu e Chu-Song Chen. «Aesthetic critiques generation for photos». In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 3514–3523 (cit. a p. 6).
- [4] IL PASSE-PARTOUT DEI BELLI. «6-Psicologia della bellezza». In: () (cit. a p. 7).
- [5] Jia Deng et al. «Imagenet: A large-scale hierarchical image database». In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. a p. 14).
- [6] Geoffrey Hinton, Oriol Vinyals e Jeff Dean. «Distilling the knowledge in a neural network». In: *arXiv preprint arXiv:1503.02531* (2015) (cit. a p. 15).
- [7] Wei-Chih Hung et al. «Learning to blend photos». In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 70–86 (cit. a p. 6).
- [8] Saira Kanwal, Muhammad Uzair e Habib Ullah. «A Survey of Hand Crafted and Deep Learning Methods for Image Aesthetic Assessment». In: *arXiv preprint arXiv:2103.11616* (2021) (cit. a p. 3).
- [9] Shu Kong et al. «Photo aesthetics ranking network with attributes and content adaptation». In: *European Conference on Computer Vision*. Springer. 2016, pp. 662–679 (cit. a p. 6).

-
- [10] Kevin S LaBar et al. «Hunger selectively modulates corticolimbic activation to food stimuli in humans.» In: *Behavioral neuroscience* 115.2 (2001), p. 493 (cit. a p. 9).
 - [11] Wei Luo, Xiaogang Wang e Xiaoou Tang. «Content-based photo quality assessment». In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 2206–2213 (cit. a p. 6).
 - [13] Charles Michel et al. «A taste of Kandinsky: assessing the influence of the artistic visual presentation of food on the dining experience». In: *Flavour* 3.1 (2014), pp. 1–11 (cit. a p. 9).
 - [14] Veranika Mikhailava, Evgeny Pyshkin e Vitaly Klyuev. «Aesthetic Evaluation of Food Plate Images using Deep Learning». In: *2020 22nd International Conference on Advanced Communication Technology (ICAET)*. IEEE. 2020, pp. 285–289 (cit. a p. 8).
 - [15] Naila Murray, Luca Marchesotti e Florent Perronnin. «AVA: A large-scale database for aesthetic visual analysis». In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 2408–2415 (cit. a p. 6).
 - [16] Yilang Peng e JOHN B JEMMOTT III. «Feast for the Eyes: Effects of Food Perceptions and Computer Vision Features on Food Photo Popularity.» In: *International Journal of Communication (19328036)* 12 (2018) (cit. a p. 9).
 - [17] Ramprasaath R Selvaraju et al. «Grad-cam: Visual explanations from deep networks via gradient-based localization». In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626 (cit. alle pp. 33, 39).
 - [18] Kekai Sheng et al. «Attention-based multi-patch aggregation for image aesthetic assessment». In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 879–886 (cit. a p. 15).
 - [19] Kekai Sheng et al. «Learning to assess visual aesthetics of food images». In: *Computational Visual Media* 7.1 (2021), pp. 139–152 (cit. alle pp. 6, 10, 13–16, 18, 30, 31, 37, 44).
 - [20] Dimitris Spathis. «Photo-Quality Evaluation based on Computational Aesthetics: Review of Feature Extraction Techniques». In: *arXiv preprint arXiv:1612.06259* (2016) (cit. a p. 29).
 - [21] Charles Spence. «On the psychological impact of food colour». In: *Flavour* 4.1 (2015), pp. 1–16 (cit. a p. 9).
 - [22] Rongju Sun et al. «Aesthetic Visual Quality Evaluation of Chinese Handwritings.» In: *IJCAI*. Vol. 15. 2015, pp. 2510–2516 (cit. a p. 6).

- [23] Christian Szegedy et al. «Rethinking the inception architecture for computer vision». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826 (cit. a p. 15).
- [24] Carlos Velasco et al. «Hedonic mediation of the crossmodal correspondence between taste and shape». In: *Food Quality and Preference* 41 (2015), pp. 151–158 (cit. a p. 9).
- [25] Susan C Wooley e Orland W Wooley. «Salivation to the sight and thought of food: a new measure of appetite.» In: *Psychosomatic Medicine* (1973) (cit. a p. 9).
- [26] Wenhui Yu et al. «Aesthetic-based clothing recommendation». In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 649–658 (cit. a p. 6).

Siti

- [12] *MATLAB*. URL: <https://it.mathworks.com/products/matlab.html> (cit. alle pp. 1, 29, 44).