Journal of Modeling in Engineering Volume Nomber (year) first page-last page
This is an Open Access paper licensed under the Creative Commons License CC BY 4.0 license.

Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/



Research Article

Recognition of Facial and Voice Emotional States Using Deep-BEL Model

Sara Motamed [a,*], Elham Askari [b]

1. Assistant Professor, Department of Computer Engineering, Fouman and Shaft Branch, Islamic Azad University, Fouman, Iran
2. Assistant Professor, Department of Computer Engineering, Fouman and Shaft Branch, Islamic Azad University, Fouman, Iran

*Corresponding Author: Sara Motamed

PAPER INFO

Paper history:

Received:
Revised:
Accepted:

Keywords:
Emotion recognition,
Facial and voice
Emotional states,
Convolutional neural networks,
Brain emotional learning model.

ABSTRACT

In recent years, emotion recognition has attracted significant attention as a key component for achieving natural human–computer interaction. Relying on a single modality (speech or facial images) often leads to ambiguities; therefore, multimodal fusion can provide more accurate performance. In this paper, a hybrid approach is proposed for emotion recognition based on emotional speech, visible facial expressions, and infrared facial images. First, deep learning models are employed to extract rich representations from both auditory and visual data. Then, the Brain Emotional Learning (BEL) model, inspired by the limbic system, is utilized to fuse the three modalities. The proposed method is evaluated on the Enterface'05 audiovisual database. Experimental results show that the model achieves a recognition accuracy of 96.20%, which significantly outperforms other fusion-based approaches on the same dataset. These findings demonstrate the efficiency and strong generalization capability of the Deep-BEL model for applications in human–computer interaction.

© 2023

UNCORRECTED PROOF

How to cite this article:

بازشناسی حالات هیجانی چهره و صدا بااستفاده از مدل Deep-BEL

سارا معتمد [۱،*]، الهام عسکری [۲]

اطلاعات مقاله

نوع مقاله:
دریافت مقاله:
بازنگری مقاله:
پذیرش مقاله:

واژگان کلیدی:
بازشناسی هیجان، حالات هیجانی چهره و صدا، شبکه‌های عصبی کانولوشنی، مدل یادگیری هیجانی مغز، سیستم لیمبیک مغز.

چکیده

در سال‌های اخیر، بازشناسی هیجان به عنوان یکی از مؤلفه‌های کلیدی در تعامل طبیعی انسان و کامپیوتر مورد توجه بسیاری از پژوهشگران قرار گرفته است. بازشناسی مبتنی بر تنها یکی از مدالیته‌ها (گفتار یا تصویر چهره) معمولاً با ابهامات همراه است، ازاین‌رو همجوشی اطلاعات چندحسی می‌تواند عملکرد دقیق‌تری ارائه دهد. در این مقاله روشی ترکیبی برای بازشناسی هیجان بر اساس گفتار هیجانی، تصاویر مرئی حالات چهره و تصاویر مادون قرمز ارائه می‌شود. در گام نخست، از شبکه‌های عمیق برای استخراج بازنمایی‌های غنی از داده‌های شنیداری و دیداری استفاده شده و سپس مدل یادگیری هیجانی مغز(BEL) ، الهام‌گرفته از سیستم لیمبیک، برای همجوشی سه مدالیته به‌کار گرفته شده است. روش پیشنهادی بر روی پایگاه داده چندرسانه‌ای Enterface’05 ارزیابی شد و نتایج نشان داد که مدل ارائه‌شده به دقت ۹۶.۲۰٪ دست یافت که نسبت به سایر روش‌های همجوشی در همین پایگاه داده بهبود قابل توجهی دارد. این نتایج نشان‌دهنده کارایی و قابلیت تعمیم بالای مدل Deep-BEL در کاربردهای تعامل انسان–کامپیوتر است.

UNCORRECTED PROOF

۱-مقدمه [۱]

ارتباط میان انسان‌ها به دو صورت کلامی و غیرکلامی برقرار می‌شود. منظور از ارتباط کلامی، جملات و واژگانی است که انسان‌ها در تعاملات روزمره به کار می‌برند. ارتباط غیرکلامی شامل فعالیت‌هایی همچون حرکات سر و بدن و تغییر لحن گفتار است که معنای انتقال‌یافته از طریق کلام را توسعه داده و تکمیل می‌کند. بر اساس نظر مهرابیان، بیش از ۹۰ درصد مفاهیم ردوبدل‌شده در ارتباطات انسانی از طریق کانال‌های غیرکلامی منتقل می‌شود و در این میان، حالات چهره، لحن گفتار و حرکات بدن نقش ویژه‌ای دارند [۱]. هیجان، بخشی مهم از ارتباط غیرکلامی محسوب می‌شود؛ به‌طور مثال، اخم در هنگام صحبت نشان‌دهنده عدم رضایت و مخالفت است، در حالی که تعجب، خوشحالی و ترس واکنش‌های طبیعی انسان‌ها در برابر محرک‌های محیطی هستند. همچنین بروز ناخودآگاه برخی احساسات در مراودات انسانی نقش مهم و جهت‌دهنده‌ای دارد. بنابراین، به‌کارگیری کامپیوتر در چنین کاربردهایی نیازمند توانایی شناسایی صحیح حالات و هیجان‌های کاربر است و سیستم باید بتواند هیجان‌ها را که در قالب لحن گفتار، حالات چهره یا حرکات بدن بروز می‌کنند، به‌درستی تشخیص دهد.

مطالعات روان‌شناسی متعددی به همجوشی فرایند شنیداری–دیداری توسط انسان پرداخته‌اند؛ برای نمونه در لب‌خوانی [2] و مکان‌یابی منبع صدا [3] نشان داده شده که انسان‌ها برای درک بهتر اطلاعات چندمدالیته را به‌طور طبیعی ترکیب می‌کنند. این مطالعات، مبنای توسعه سیستم‌های هوشمند برای ترکیب مدالیته‌های شنیداری و دیداری فراهم کرده‌اند. با این وجود، پیشرفت‌های بیشتری لازم است تا کامپیوتر بتواند در سطح انسان اطلاعات چندمدالیته را پردازش کند. بررسی‌های گذشته نیز ضرورت تحقیقات بیشتر در زمینه همجوشی اطلاعات را تأیید کرده‌اند [۴–۶].

بر این اساس، در این مقاله از همجوشی اطلاعات شنیداری–دیداری به‌منظور بازشناسی حالات هیجانی استفاده شده است. سلسله مراحل مدل پیشنهادی بدین صورت است: در گام نخست، سه ورودی شامل تصاویر مرئی چهره، سیگنال گفتار، و تصاویر مادون قرمز حاصل از پردازش تصاویر مرئی با استفاده از نرم‌افزار پایتون از پایگاه داده Enterface’05 استخراج می‌شوند. در مرحله دوم، تمامی ورودی‌ها تحت پیش‌پردازش قرار می‌گیرند. سپس در مرحله سوم، از تصاویر مرئی، تصاویر مادون قرمز و سیگنال‌های گفتار ویژگی‌های مناسب استخراج شده و در ادامه، بهترین ویژگی‌ها با یکدیگر همجوشی می‌شوند. بردارهای به‌دست‌آمده به بخش‌های آمیگدال و اوربیتوفرانتال مدل یادگیری هیجانی مغز ارسال می‌شوند تا نرخ بازشناسی حالات هیجانی مبتنی بر حالات چهره و گفتار محاسبه شود. دلیل انتخاب این سه ورودی (تصاویر مرئی، تصاویر مادون قرمز و سیگنال گفتار) آن است که امکان استخراج دقیق‌تر ویژگی‌ها فراهم شود و علاوه بر این، تصاویر مادون قرمز قادر به آشکارسازی حالات درونی همچون استرس یا اضطراب هستند که ممکن است در تصاویر مرئی چهره به‌طور کامل نمایان نشوند. نوآوری‌های این مقاله به شرح ذیل است:

- ارائه مدل **Deep-BEL** به‌عنوان یک چارچوب ترکیبی جدید که شبکه‌های عصبی کانولوشنی (CNN) را با مدل یادگیری هیجانی مغز (BEL) –الهام گرفته از سیستم لیمبیک – ادغام می‌کند. به‌کارگیری CNN در بخش تالاموس و قشر حسی مدل BEL به‌منظور استخراج عمیق ویژگی‌ها از داده‌های چندمدالیته (گفتار، تصاویر مرئی چهره و تصاویر مادون قرمز).

* پست الکترونیک نویسنده مسئول: motamed.sarah@gmail.com
۱. استادیار، گروه کامپیوتر، واحد فومن و شفت، دانشگاه آزاد اسلامی، فومن، ایران
۲. استادیار، گروه کامپیوتر، واحد فومن و شفت، دانشگاه آزاد اسلامی، فومن، ایران

UNCORRECTED PROOF

طراحی راهبرد همجوشی چندمدالیته با استفاده از تحلیل همبستگی کانونی (CCA) برای ترکیب ویژگی‌های استخراج‌شده و انتقال آن‌ها به بخش‌های آمیگدال و اوربیتوفرانتال مدل BEL جهت طبقه‌بندی نهایی. دستیابی به دقت ۹۶.۲۰٪ بر روی پایگاه داده Enterface'05که عملکرد مدل را نسبت به سایر روش‌های همجوشی ارتقا داده و برتری روش پیشنهادی را نشان می‌دهد.

بخش‌های اصلی این مقاله به این صورت سازماندهی شده است که بخش ۲ مربوط به بررسی کارهای انجام شده روی بازشناسی حالات هیجانی چهره، صدا و همجوشی آنها می‌باشد. در بخش ۳ روش پیشنهادی DEEP-BEL ارائه خواهد شد و نتایج آزمایشات و نتیجه‌گیری در بخش‌های ۴ و ۵ مطرح می‌شوند.

۲- بازشناسی حالات هیجانی

۲-۱-بازشناسی هیجان از روی چهره

هر بیان احساس در ارتباطات روزمره از جایگاه ویژه‌ای برخوردار است. یکی از مهمترین بسترهای مهم هیجان چهره است. از این‌رو یکی از جنبه‌های مهم در تعامل طبیعی میان انسان و ماشین، بازشناسی حالات چهره و تولید بازخورد متناسب با هیجان درک شده می‌باشد. این تغییرات ترکیبی از حرکت ماهیچه‌هایی است که در نواحی مختلف چهره قرار دارند. در مقاله [۷]، تغییرات چهره را در قالب فعالیت‌ها روی ماهیچه‌ها بررسی کرده‌اند و تعدادی از آنها را در قالب سیستم کدگذاری حرکات چهره (FACS[2]) گردآوری شدند. در این سیستم، واحد حرکت[3]، به تغییری در چهره اطلاق می‌شود که اولاً به تنهایی انجام پذیرد و ثانیاً غیرقابل تقسیم باشد. در [۸] از مدل شکلی[4] که بوسیله ۵۸ نقطه برجسته در چهره تعریف شده است استفاده کردند. در [۹]، مجموعه‌ای از نقاط شاخص اطراف لب، چشم‌ها، ابروها، بینی و چانه را مورد بررسی قرار دادند. ویژگی‌های ظاهری نیز از طریق روش‌هایی مانند موجک گابور[5] [۱۰]، ویژگی Haar [۱۱]، قالب چهره نسبت فاصله‌ای[6] [۱۲] و غیره بدست آمد. در بسیاری از مطالعات [۱۳]، استفاده از ویژگی‌های ظاهری و هندسی، بهترین گزینه برای طراحی بازشناسی حالات چهره می‌باشد. در این میان، بازنمایی Local Binary Patterns [۱۴] و Local Phase Quantization [۱۵] دو بازنمایی از روش‌های استخراج ویژگی مبتنی بر ظاهر می‌باشند. در سال ۲۰۰۵، Sebe، کارایی سیستم‌های موجود بین ۷۴٪ و ۹۸٪ درصد را بسته به نوع الگوریتم، تعداد دسته‌های هیجان و غیره بررسی کرد. جزئیات بیشتری در تحقیق [۱۶] انجام شده است. در سال‌های اخیر، روش‌های مبتنی بر یادگیری عمیق پیشرفت چشمگیری در بازشناسی حالات چهره داشته‌اند. برای مثال، در[۳۰]، با استفاده از همجوشی مبتنی بر توجه بر روی داده‌های چندرسانه‌ای، عملکرد بالاتری نسبت به روش‌های کلاسیک به دست آوردند. همچنین مطالعات دیگری همچون [۳۱] و [۳۲] با به‌کارگیری معماری‌های LSTM وTransformer، نتایج قابل توجهی در همجوشی اطلاعات دیداری–شنیداری گزارش کرده‌اند.

۲-۲- بازشناسی هیجان از روی گفتار

سیگنال گفتار سریعترین و طبیعی‌ترین روش برای نمایش ارتباط متقابل میان انسان‌هاست. این واقعیت محققان را بر آن داشته است تا از گفتار به عنوان روشی سریع در ارتباط میان ماشین و انسان استفاده کنند [۱۷،۱۸]. امروزه تحقیقات بسیاری در زمینه‌ی پردازش سیگنال‌های گفتار و طبقه‌بندی آنها بمنظور بازشناسی گفتار هیجانی مبتنی بر مدل‌های یادگیری ماشین انجام شده است. برای مثال شبکه عصبی [۱۹]، ماشین بردار پشتیبان[۲۰]، k نزدیکترین همسایگی [۲۱]، مدل مخفی مارکوف [۲۲]، مدل مخلوط گوسی [۲۳] و غیره. برای مثال در [۲۴] از مدل شبکه عصبی چند لایه پرسپترون و هفت نوع هیجان استفاده کرده‌اند و بالاترین نرخ بازشناسی به

[2] Facial Action Coding System

[3] Action Units

[4] Shape Model

[5] Gabor Wavelets

[6] Spatial Ratio Face Template

UNCORRECTED PROOF

دست آمده در آزمایشات آنها ۵۳٪ بوده است. همچنین در [۲۵] از مدل‌های طبقه‌بندی ماشین بردار پشتیبان و شبکه عصبی و چهار گونه هیجان خشم، طبیعی، خوشحالی و ترس استفاده کرده و نرخ بازشناسی در آزمایشات آنها برای مدل‌های ماشین بردار پشتیبان و شبکه عصبی مصنوعی به ترتیب ۷۱٪ و ۴۲٪ بوده است. در مقاله [۲۶] از شبکه‌های عصبی کانولوشنی (CNN) بمنظور استخراج ویژگی گفتار استفاده کرده است. در لایه اول CNN و بمنظور تشخیص بازشناسی گفتار و جهت استخراج ویژگی‌های سیگنال‌های هیجانی کرده و اخیرا از CNN دو لایه و LSTM بمنظور بازشناسی گفتار هیجانی end-to-end استفاده شده است. در [۲۷] DCNNs که متشکل از لایه‌های کانولوشنی عمیق چند سطحی و لایه‌های Pooling است، استفاده شدند. از آنجا که DCNNs شامل پارامترهای بیشتری برای استخراج فرکانس زمانی است، دقیق‌تر بوده. همچنین اطلاعات همبستگی و توانایی یادگیری با ویژگی قوی آنها موجب عملکرد بهتر در CNNs کم عمق شده است. در [۲۸] از DCNN, AlexNet, برای بازشناسی گفتار هیجانی استفاده شده است. در [۳۳] با توسعه WavFusion بر پایه wav2vec 2.0 توانستند در بازشناسی گفتار هیجانی بهبود قابل توجهی ایجاد کنند.

۲-۳- بازشناسی هیجان بااستفاده از همجوشی اطلاعات در سطح ویژگی

طبق بررسی‌های انجام شده، تحقیقات زیادی در زمینه بازشناسی هیجان با استفاده از گفتار یا حالات چهره انجام شده است. اما تلاش‌های معدودی برای ترکیب این دو صورت گرفته است. ترکیب دو رسانهٔ، گفتار و چهره، زمینه تحقیقات جدیدی را برای ارتباط چندرسانه‌ای انسان و ماشین فراهم می‌کند. از مهمترین چالش‌های این مسئله این است که در حالت طبیعی بروز احساس، ویژگی‌های چهره و گفتار، همزمان نمی‌باشند. اما از آنجا که این طرح بر روی همجوشی در سطح بالای ویژگی متمرکز می‌شود، فرض شده ویژگی‌های هیجان گفتار و حالات چهره همزمان می‌باشند. برای همزمان‌سازی رسانه‌های مختلف، روش‌های مختلفی ارائه شده است. در[۲۹]، از بافرهای فعالی برای نگه‌داری مقادیر ویژگی‌ها و همجوشی زمانی آنها استفاده کرده‌اند. محققان علوم شناختی به این همجوشی ویژگی‌ها به عنوان استراتژی همجوشی اولیه اشاره می‌کنند. مزیت همجوشی اولیه این است که قابلیت بهتری در جداسازی برای طبقه‌بندها با استفاده از کوواریانس بین ویژگی‌های شنیداری و دیداری فراهم شده است [۳۰]. اخیراً، پژوهش‌های جدیدتر از همجوشی سطح ویژگی فراتر رفته و به سمت معماری‌های پیشرفته‌تر رفته‌اند. برای نمونه، در[۳۴] با ترکیب چند مدالیته شامل چهره، گفتار و حرکات بدن، عملکرد بهتری نسبت به روش‌های فیوژن سنتی ارائه داده‌اند. این مطالعات نشان می‌دهد که همجوشی عمیق مبتنی بر یادگیری توجه و معماری‌های پیشرفته، مسیر اصلی تحقیقات اخیر است.

۳- مدل پیشنهادی

دراین بخش ما ابتدا مدل پیشنهادی و سپس زیر ساختار هر جزء از مدل پیشنهادی را توضیح خواهیم داد:

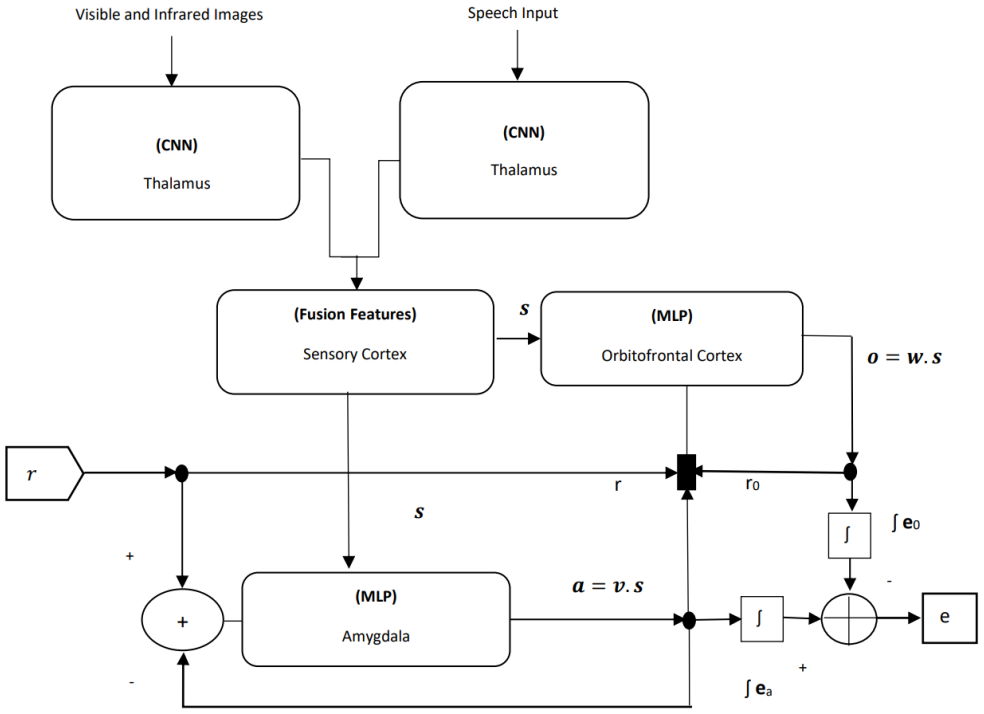
۳-۱- مدل یادگیری هیجانی مغز کانولوشنی (DEEP-BEL)

پس ورودی مدل پیشنهادی DEEP-BEL ، تصاویر حالات مختلف هیجانی چهره واقعی و مادون قرمز و سیگنال‌های گفتار می‌باشند. ورودی‌ها پس از خواندن نرمالسازی شده و وارد مدل پیشنهادی خواهند شد. دلیل انتخاب مدل BEL این است که این مدل الهام گرفته از سیستم بیولوژیکی لیمبیک مغز انسان است. البته در این مدل تغییراتی اعمال شده و از شبکه عصبی کانولوشنی در بخش تالاموس استفاده شده است. از مزایای شبکه عصبی کانولوشنی این است که برای دستیابی به نتایج بهتر و دقیق می‌توان شبکه را تنظیم کرد. هر چقدر این شبکه بهتر تنظیم شود و مقدار کافی از داده‌ها را تغذیه کند، می‌تواند نتایج بهتری نسبت به الگوریتم‌های یادگیری ماشین ارائه دهد. هر نورون به همه‌ی سلول‌های عصبی موجود در لایه قبلی متصل نیست، بلکه فقط به تعداد کمی نورون متصل است. این مسئله باعث کاهش بسیاری از پارامترها می‌شود. از طرفی دیگر مجموعه‌ای از

UNCORRECTED PROOF

اتصالات می‌توانند وزن یکسانی داشته باشند، به همین دلیل باعث کاهش بسیاری از پارامترها می‌شود. در انتها هم می‌توان به این نکته اشاره کرد که لایه Pooling می‌تواند با حفظ اطلاعات مفید، مقدار پردازش داده را کاهش دهد و برای دیتاست‌ها با اندازه بزرگ مناسب می‌باشد. از آنجاییکه شبکه‌های عصبی کانولوشنی مزایای زیادی دارد، در این مقاله بعنوان استخراجگر ویژگی از این شبکه استفاده شده است.

در انتها بردارهای ویژگی استخراج شده به شبکه‌های MLP در آمیگدال و اوربیتوفرانتال، در مدل DEEP-BEL ارسال می‌شود. همانطور که در شکل (۱) نشان داده شده، این مدل براساس پاداش و جریمه کار می‌کند و با r و r_0 نمایش داده شده است.



شکل ۱ – شمای کلی روش پیشنهادی

سلسله مراحل مدل پیشنهادی در شکل (۱) نمایش داده شده است. همانطور که در شکل (۱) نشان داده می‌شود، اولین مرحله خواندن ویدئوها از دیتاست Enterface'05 می‌باشد [۳۱]. برای جداسازی سیگنال گفتار و ویدئو الگوریتمی با جداسازی نرخ فریم مشخص روی همه نمونه‌ها اعمال خواهد شد. سپس تمامی تصاویر واقعی و مادون قرمز از چهره و سیگنال‌های گفتار در پوشه‌های مشخص ذخیره می‌شوند. در ادامه استخراج ویژگی در ناحیه تالاموس از مدل DEEP-BEL انجام خواهد شد. تمامی ویژگی‌های استخراج شده همجوشی شده و خروجی به آمیگدال و اوربیتوفرانتال انتقال داده خواهد شد تا کلاسبندی روی این ویژگی‌ها انجام شود. خروجی مدل پیشنهادی تشخیص حالات هیجانی خواهد بود.

۳-۱-۱- دیتاست Enterface’05

پایگاه داده گفتاری- تصویری Enterface'05 از ۴۴ نفر تشکیل شده است. که هر نفر ۶ هیجان (خشم، انزجار، ترس، شادی، غم، تعجب) پایه را با ۵ سناریو و جملات مختلف بیان می‌کنند. طول هر ویدئو بطور متوسط ۳-۴ ثانیه می‌باشد و زبان این پایگاه داده انگلیسی است. در کل پایگاه داده شامل ۱۳۲۰ نمونه کلیپ گفتاری-تصویری است که سیگنال گفتار آن با

UNCORRECTED

نرخ نمونه برداری ۴۸۰۰۰ هرتز با رزلوشن ۱۶ بیت و یک کاناله ضبط شده است.

۳-۱-۲- پیش پردازش

برای بازشناسی هیجان از روی حالات چهره در ویدئو در مرحله اول آماده‌سازی پایگاه داده و عملیات پیش پردازش انجام می‌شود. این مراحل بدینصورت است: در مرحله آماده‌سازی پایگاده داده ابتدا، مجموعه پایگاه داده به ۲ بخش آموزش و آزمایش تقسیم می‌شود. بطوری که ۷۰٪ مجموعه پایگاه داده برای آموزش و ۳۰٪ آن برای آزمایش در نظر گرفته می‌شود. در مرحله بعد، فریم‌های تصاویر چهره و سیگنال گفتار از هر نمونه ویدئو استخراج شده و در پوشه‌های مربوط به فریم ویدئو و گفتار به ترتیب ذخیره می‌شوند. برای استخراج فریم‌های تصویر و سیگنال گفتار از ویدئو الگوریتم جداسازی با نرخ فریم مشخص بر روی همه نمونه‌ها اعمال خواهد شد. در این آزمایشات نرخ فریم تصاویر ۱۰ (یعنی در هر ثانیه ۱۰ فریم را استخراج می‌کند) می‌باشد.

۳-۱-۳- استخراج ویژگی

در روش پیشنهادی، از شبکه عصبی مصنوعی برای یادگیری عمیق و طبقه‌بندی احساسات استفاده شده است. این مدل از دو لایه تشخیص تشکیل شده و در لایه اول سه CNNs روی تصاویر مرئی و مادون قرمز و سیگنال‌های گفتار به صورت انفرادی آموزش خواهند دید. سپس ویژگی‌های بدست آمده همجوشی شده و خروجی ویژگی‌های همجوشی شده به بخشهای آمیگدال و اوربیتوفرانتال بمنظور طبقه‌بندی انتقال می‌یابند. طبقه‌بندی نهایی به عنوان خروجی لایه دوم بدست آمده است.

۳-۱-٤- همجوشی اطلاعات

در این مقاله معماری مدل پیشنهادی بازشناسی هیجان مبتنی بر همجوشی دو فرآیند شنیداری-دیداری از دو مسیر متفاوت بدست می‌آیند. یک جریان، پردازش بینایی است که فریم‌های تصاویر واقعی و مادون قرمز از چهره را پس از پیش‌پردازش در قسمت تالاموس وارد قشر بینایی مغز کرده و از طریق شبکه عصبی کانولوشن، اطلاعات تصاویر چهره را بدست می‌آورد. جریان دیگر، پردازش شنوایی است که پس از عملیات پیش پردازش مانند فریم‌بندی و پنجره‌زنی در قسمت تالاموس، سیگنال گفتار به تصویر مل اسپکتروگرام تبدیل شده و به قشر شنوایی برای پردازش‌های پیچیده‌تر ارسال می‌شوند. در قشر شنوایی از CNN بمنظور بازنمایی ویژگی‌ها استفاده می‌گردد.

در این مرحله AlexNet استفاده شده و شبکه بدست آمده شامل ۲۵ لایه عمیق CNN است. در لایه کانولوشنال از فیلترهای کانولوشنی متحرک در تصاویر ورودی استفاده می‌گردد و خروجی ایجاد شده ویژگی‌های نگاشت شده خواهند بود. چند لایه واحد خطی اصلاح شده (ReLU) در CNN وجود دارد. الکسی نت برای طبقه‌بندی احساسات با جایگزینی سه لایه آخر با لایه‌های جدید که برای یادگیری از پایگاه داده جدید منصوب شدند، تعریف خواهد شد. ۵ لایه اول (دو لایه تماما متصل، یک فعال ساز، یک سافت مکس و یک لایه کلاسبندی) استخراج شده و برای مرحله همجوشی در سطح ویژگی استفاده می‌شوند. ما همچنین لایه‌های جمع‌آوری را که عملکرد اولیه آن کاهش اندازه‌ی محاسبات انجام شده است و در لایه‌های بعدی به کاهش اتصالات بیش از حد و کاهش سربار کمک می‌کند، استفاده کرده‌ایم. همچنین از چندین لایه‌های dropouts، کاملاً متصل، و نرمال‌سازی کانال نیز استفاده شده است. لایه‌های نهایی شامل یک سافت ماکس (که تابع سافت ماکس را اعمال می‌کند) و یک لایه طبقه‌بندی است که به عنوان ورودی احتمالی کلاس داده می‌شود. در این بخش و بمنظور همجوشی در سطح ویژگی از روش همبستگی کانونی (CCA[7]) [۳۲]، جهت ترکیب ویژگی‌های استخراج شده از دو CNNs در لایه اول چارچوب پیشنهادی استفاده شده است. CCA یک روش تحلیل آماری چند متغیره برای ارتباط متقابل بین چند بردار دلخواه است. از طرف دیگر سیگنال‌های گفتار پس از خواندن نرمالسازی می‌شوند و مل

[۷] Canonical Correlation Analysis

UNCORRECTED PROOF

اسپکتروگرام بمنظور تبدیل سیگنال‌ها به تصاویر تبدیل خواهند شد. تصاویر مل اسپکتروگرام گفتار با اندازه ۱×۲۰۰×۹۶ به شبکه عصبی CNN داده شده و پس از اعمال لایه‌های مختلف یک بردار ویژگی ۲۵۶ تایی از گفتار بدست می‌آید. ویژگی‌های همجوشی شده از تصاویر واقعی و مادون قرمز با ویژگی‌های بدست آمده از گفتار با هم همجوشی خواهند شد. ما از روش جمع‌بندی برای ترکیب بردارهای ویژگی تبدیل شده استفاده کردیم. ویژگی‌های همجوشی شده ورودی مدل کلاسبندی پیشنهادی خواهند بود. جهت توضیح اضافه، ویژگی‌های الحاق شده بطور موازی به آمیگدال و اوربیتوفرانتال در مدل DEEP-BEL داده می‌شوند. این شبکه پس از همجوشی ویژگی‌ها، خروجی کلاس نهایی هیجان را می‌دهد.

۳-۲-مدل یادگیری هیجانی مغز (BEL)[۸]:

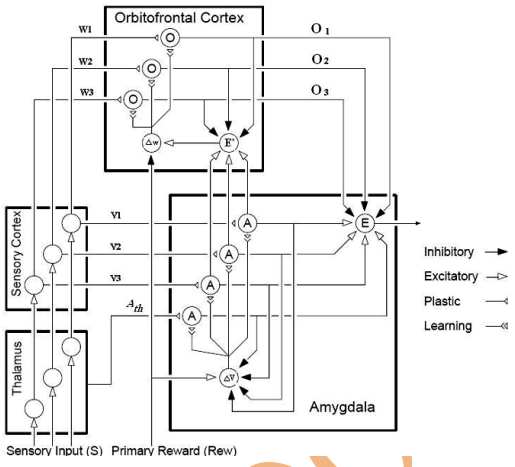
یادگیری هیجانی در مغز یک نوع از یادگیری شرطی است که با محرک‌های هیجانی خارجی مانند پاداش و جریمه دریافتی از موقعیت‌های مختلف زندگی واقعی (از جهان بیرونی) در ارتباط است. این محرک‌های هیجانی می‌تواند باعث حالات هیجانی داخلی مختلفی مانند شادی، غم و ترس شود که تصمیمات آینده ما را تحت تاثیر قرار می‌دهد. مدل‌های زیادی برای نمایش عملکرد هیجان تولیدی در مغز ارائه شده است. برای مثال مدل مورن، بل، بالکنیوس. اما جامع‌ترین مدلی که الهام گرفته از سیستم لیمبیک که مغز پستانداران است، مدل بل می باشد. اخیراً، مدل‌های مختلفی از یادگیری هیجانی مغز پستانداران (BEL) در کنترل برنامه‌های کاربردی [۳۳] و پیش‌بینی مسائل [۳۴] استفاده شده است. این مدل‌ها مبتنی بر مدل آمیگدالا و اوربیتوفرانتال می‌باشند. آمیگدلا و اوربیتوفرانتال دو قسمت اصلی از سیستم لیمبیک در مغز می‌باشند. در این مدل، آمیگدلا یاد می‌گیرد که به محرک هیجانی پاسخ دهد و اوربیتوفرانتال تجربه‌های نامناسب و اتصالات یادگیری را مهار کند. مدل‌های یادگیری می‌بایست ویژگی سرعت بالا و پیچیدگی محاسباتی پایین را دارا باشند و مدل BEL دارای این مزایا می‌باشد. در این مقاله، پاداش و جریمه نقش مهمی را در یادگیری ایفا می‌کنند. یادگیری مدل بل بصورت تطبیقی می‌باشد یعنی این شبکه در حین آموزش، پارامترهای خود را برای رسیدن به جواب مورد نظر یاد می‌گیرد. این عمل با استفاده از تابع Rew که یک تابع هزینه از پاداش و جریمه است، بصورت تقویتی محاسبه می‌شود. آمیگدلا و اوربیتوفرانتال در مدل محاسباتی دارای ساختار شبکه‌ای هستند. بردار S ورودی محرک به سیستم را نشان می‌دهد. برای هر محرک S یک گره A وجود دارد. همانطور که در شکل (۲) مشاهده می‌شود، ورودی A_{th} از تالاموس به بخش آمیگدلا می‌رود و مقدار آن برابر با حداکثر مقدار ورودی‌های حسی می‌باشد (رابطه ۱).

$$A_{th} = \max(S_i) \qquad (۱)$$

از آنجاییکه در زیر بخش‌های تالاموس و قشر حسی از مدل BEL، روشی قوی جهت استخراج ویژگی‌ها مشخص نشده است، در این مقاله از شبکه عصبی کانولوشنی در زیر بخش‌های تالاموس و قشر حسی از مدل BEL استفاده شده است. دلیل استفاده از شبکه عصبی کانولوشنی در زیر بخش‌های تالاموس و قشر حسی از مدل BEL این است که نتایج بهتر و دقیق‌تری را در خروجی نمایش می‌دهد و با داشتن لایه Pooling، اطلاعات مفید را حفظ کرده و مقدار پردازش داده را کاهش می‌دهد. لذا برای دیتاست‌ها با اندازه بزرگ مناسب می‌باشد. به همین دلیل در مدل پیشنهادی از شبکه عصبی کانولوشنی بعنوان استخراجگر ویژگی استفاده شده است. در ادامه ویژگیهای استخراج شده به شبکه‌های MLP در آمیگدال و اوربیتوفرانتال از مدل BEL ارسال می‌شود و زیر بخش‌های مدل DEEP-BEL بطور کامل توضیح داده می‌شوند.

[۸] Brain Emotional Learning (BEL)

UNCORRECTED PROOF



شکل ۲- مدل BEL [۳۲]

۳-۲- طبقه‌بندی در مدل DEEP-BEL

در این قسمت طبقه‌بندی بااستفاده از مدل BEL صورت می‌گیرد. بردار بدست آمده از انتخاب ویژگی از مرحله قبل (S)، بطور همزمان وارد آمیگدال و اوربیتوفرانتال می‌شوند. برای هر محرک s یک گره a در مدل شبکه آمیگدال وجود دارد (شکل (۲)). یک ارتباط پلاستیکی وزنی v_i به s_i وجود دارد. خروجی هر گره با حاصل ضرب هر ورودی در وزن v توسط معادله ۲ بدست می‌آید [۳۲]. یک ارتباط وزن‌دار با وزن V بین هر گره A و ورودی حسی وجود دارد. خروجی هر گره A بوسیله ضرب ورودی در وزن V بدست می‌آید (رابطه ۲).

$$A_i = S_i V_i \qquad (۲)$$

همچنین یک ارتباط وزن‌دار با وزن W بین هر گره A و ورودی حسی وجود دارد. خروجی هر گره A بوسیله ضرب ورودی در وزن W بدست می‌آید (رابطه ۳).

$$O_i = S_i W_i \qquad (۳)$$

تغییرات وزن‌های V و W در فرآیند یادگیری به ترتیب با استفاده از روابط ۴ و ۵ انجام می‌شود:

$$\Delta W_i = \beta S_i (E' - Rew) \qquad (۴)$$

$$\Delta v_i = \alpha (S_i max\left(0, Rew - \sum_j A_j\right)) \qquad (۵)$$

که α و β نرخ یادگیری برای آمیگدلا و اوربیتوفرانتال به ترتیب می‌باشند و سیگنال تقویتی Rew یک تابع هزینه از پاداش/جریمه می‌باشد که بسته به مسئله به روش‌های مختلفی محاسبه می‌شود. نهایتاً خروجی مدل از رابطه ۶ بدست می‌آید:

$$E = \sum_i A_i - \sum_i O_i \qquad (۶)$$

همانطور که در روابط ۴ و ۵ مشاهده می‌شود، مقادیر A_i نمی‌توانند کاهش پیدا کنند. زیرا که وزن V نمی‌تواند کاهش یابد. به این دلیل است که وقتی یک واکنش هیجانی یاد گرفته می‌شود نباید فراموش شود. یعنی فراموش کردن اطلاعات یاد گرفته شده در آمیگدلا انجام نمی‌شود. در حقیقت، "فراموش کردن" یا به تعبیری بازدارندگی وظیفه قشر اوربیتوفرانتال است. یعنی وزن W در اوربیتوفرانتال افزایش و کاهش می‌یابد.

٤- بررسی نتایج و آزمایشات

٤-١- بررسی نتایج حالات هیجانی چهره

تمامی تصاویر برای خواندن، استخراج ویژگی‌ها و در نهایت طبقه‌بندی وارد مدل DEEP-BEL می‌شوند. جدول ۱ نرخ ماتریس درهم ریختگی بازشناسی حالات چهره را روی شش حالت هیجانی نشان می‌دهد. جدول ۱ نتایج بازشناسی حالات هیجانی را بااستفاده از مدل DEEP-BEL تنها روی تصاویر چهره از دیتاست Enterface'05 نمایش می‌دهد. با توجه به ماتریس درهم ریختگی جدول ۱، بالاترین نرخ تشخیص حالات هیجانی چهره مربوط به حالت هیجانی تنفر است. همچنین پایین ترین نرخ تشخیص مربوط به حالت هیجانی ترس می‌باشد.

٤-٢- بررسی نتایج حالات هیجانی گفتار

جدول ۲ نتایج بازشناسی گفتار هیجانی را بااستفاده از مدل DEEP-BEL تنها روی سیگنال‌های گفتار از دیتاست Enterface'05 نمایش می‌دهد. با توجه به ماتریس درهم ریختگی جدول ۲، بالاترین نرخ تشخیص گفتار هیجانی مربوط به حالت هیجانی عصبانیت و با نرخ ۸۲٪ است. همچنین پایین ترین نرخ تشخیص مربوط به حالت هیجانی تنفر می‌باشد.

UNCORRECTED PROOF

٤-۳- بررسی نتایج حالات هیجانی چهره – گفتار

جدول ۳ نرخ بازشناسی از همجوشی ویژگی‌های بدست آمده روی حالات چهره و هیجان گفتار را در مدل DEEP-BEL نشان می‌دهد. با توجه به ماتریس درهم ریختگی جدول ۳، بالاترین نرخ تشخیص از همجوشی حالات هیجانی چهره و گفتار مربوط به حالت هیجانی تنفر و با نرخ ۹۹٪ و بعد از آن خوشحالی با نرخ ۸۹٪ است. همچنین پایین ترین نرخ تشخیص مربوط به حالت هیجانی عصبانیت می‌باشد. با بررسی نتایج می‌توان نتیجه گرفت که ترکیب بیومتریک‌ها نرخ بازشناسی را افزایش می‌دهد. مدل DEEP-BEL مدل بهبود یافته BEL است. همانطورکه اشاره شد این مدل الهام گرفته از سیستم لیمبیک مغز می‌باشد. وردی این مدل دیتاست Enterface'05 است. سلسه روال کار بدینصورت است که پس از خواندن تصاویر ۱۲ دنباله از تصاویر چهره به شبکه CNN داده می‌شوند و پس از اعمال لایه‌های مختلف، در لایه تماماً اتصال آخر یک بردار ویژگی ۲۵۶ تایی بدست می‌آید.

جدول ۱- ماتریس درهم ریختگی بازشناسی حالات چهره بااستفاده از مدل DEEP-BEL برای ۶ کلاس هیجان پایه بر روی پایگاه داده Enterface'05

	عصبانیت	تنفر	ترس	خوشحالی	ناراحتی	تعجب
عصبانیت	**۸۵٪**	۷.۱٪	۲.۳٪	۲.۳٪	٪۱	۲.۳٪
تنفر	۲.۳٪	**۸۸.۱٪**	۲.۳٪	۲.۳٪	۴.۷٪	۰.۳٪
ترس	۲.۳٪	۰٪	**۷۳.۸٪**	۷.۶٪	۱۴٪	۲.۳٪
خوشحالی	۷.۱٪	۳.۳٪	۴.۰٪	**۷۸.۶٪**	۲.۳٪	۴.۷٪
ناراحتی	۲.۳٪	٪۰	۹.۵٪	٪۰	**۸۱٪**	۷.۲٪
تعجب	۲.۳٪	٪۰	۴.۷٪	۷.۱٪	۷.۱٪	**۷۸.۸٪**

جدول ۲ – ماتریس درهم‌ریختگی بازشناسی هیجان گفتار بااستفاده از مدل DEEP-BEL برای ۶ کلاس هیجان پایه بر روی پایگاه داده Enterface'05

	عصبانیت	تنفر	ترس	خوشحالی	ناراحتی	تعجب
عصبانیت	**٪۸۲**	٪۷	٪۰	٪۳	٪۳	٪۵
تنفر	٪۱۲	**٪۵۵**	٪۱۰	٪۷	٪۷	٪۹
ترس	٪۲	٪۵	**٪۶۰**	٪۱۴	٪۱۴	٪۵

UNCORRECTED PROOF

خوشحالی	٪۱۲	٪۲	٪۵	**٪۶۹**	٪۲	٪۱۰
ناراحتی	٪۵	٪۵	٪۱۰	٪۰	**٪۷۰**	٪۱۰
تعجب	٪۵	٪۱۰	٪۲	٪۱۵	٪۵	**٪۶۳**

همچنین سیگنالهای گفتار پس از اعمال مل-اسپکتروگرام به شبکه عصبی CNN داده شده و پس از اعمال لایه‌های مختلف یک بردار ویژگی ۲۵۶ تایی از گفتار بدست می‌آید. سپس این ویژگی‌ها الحاق شده و به آمیگدال و اوربیتوفرانتال ارسال شده و روی آن شبکه MLP اعمال می‌شود. در نهایت خروجی بازشناسی ۶ حالت هیجانی مبتنی بر همجوشی حالات هیجانی چهره و گفتار است. جدول ۴ مقایسه‌ای از روش‌های اخیر بازشناسی هیجان چندمدالیته با روش پیشنهادی ارائه می‌دهد. همان‌طور که مشاهده می‌شود، پژوهش‌های اخیر [۳۰–۳۴] اغلب بر روی ترکیب مدالیته‌های گفتار و چهره تمرکز داشته و از ساختارهای یادگیری عمیق متنوعی همچون CNN، LSTM و Transformer بهره گرفته‌اند. اگرچه این روش‌ها دقت بالایی دارند، اما معمولاً یا محدود به دو مدالیته هستند و یا فاقد داده‌های مادون قرمز می‌باشند. روش پیشنهادی Deep-BEL با بهره‌گیری همزمان از سه مدالیته (گفتار، تصویر مرئی و مادون قرمز چهره) و ترکیب شبکه‌های کانولوشنی با مدل BEL الهام‌گرفته از سیستم لیمبیک، توانسته است دقت ٪۹۶.۲۰ را بر روی پایگاه داده Enterface’05 کسب کند. این نتایج نشان می‌دهد که مدل پیشنهادی علاوه بر ارائه عملکرد رقابتی با روش‌های به‌روز، از مزیت نوآوری در استفاده از مدالیته مادون قرمز و ساختار بیولوژیکی نیز برخوردار است.

جدول ۳ – ماتریس درهم‌ریختگی بازشناسی حالات هیجانی چهره – گفتار بااستفاده از مدل DEEP-BEL برای ۶ کلاس هیجان پایه بر روی پایگاه داده Enterface'05

	عصبانیت	**تنفر**	**ترس**	**خوشحالی**	**ناراحتی**	**تعجب**
عصبانیت	**٪۸۴**	٪۲	٪۷	٪۱	٪۴	٪۲
تنفر	٪۰	**٪۹۹**	٪۱	٪۰	٪۰	٪۰
ترس	٪۱	٪۴	**٪۸۹**	٪۰	٪۶	٪۰
خوشحالی	٪۰	٪۱	٪۰	**٪۹۷**	٪۱	٪۱
ناراحتی	٪۰	٪۲	٪۱	٪۲	**٪۹۵**	٪۰
تعجب	٪۰	٪۲	٪۶	٪۱	٪۰	**٪۹۱**

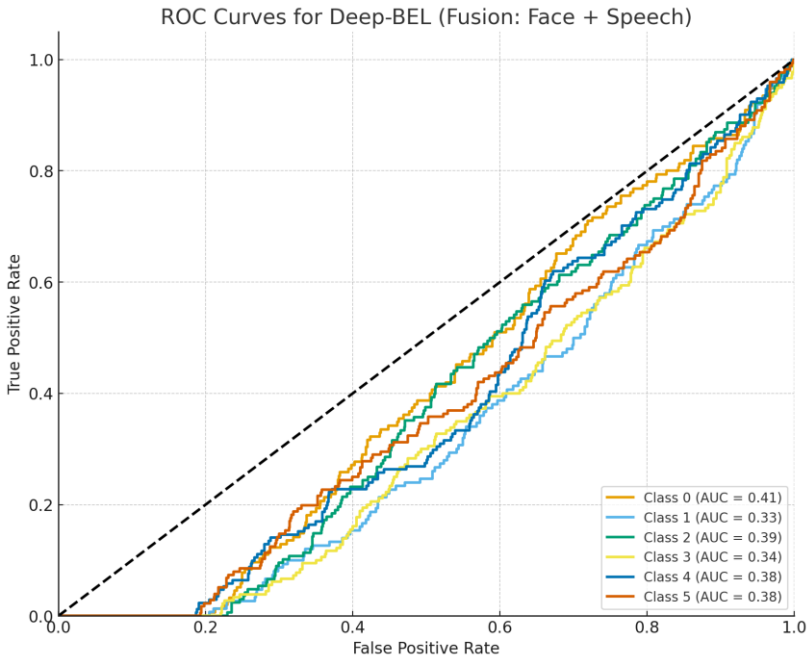
UNCORRECTED PROOF

جدول ۴ – مقایسه دقت بازشناسی مدل‌های پیشنهادی به ازای حالات مختلف هیجان

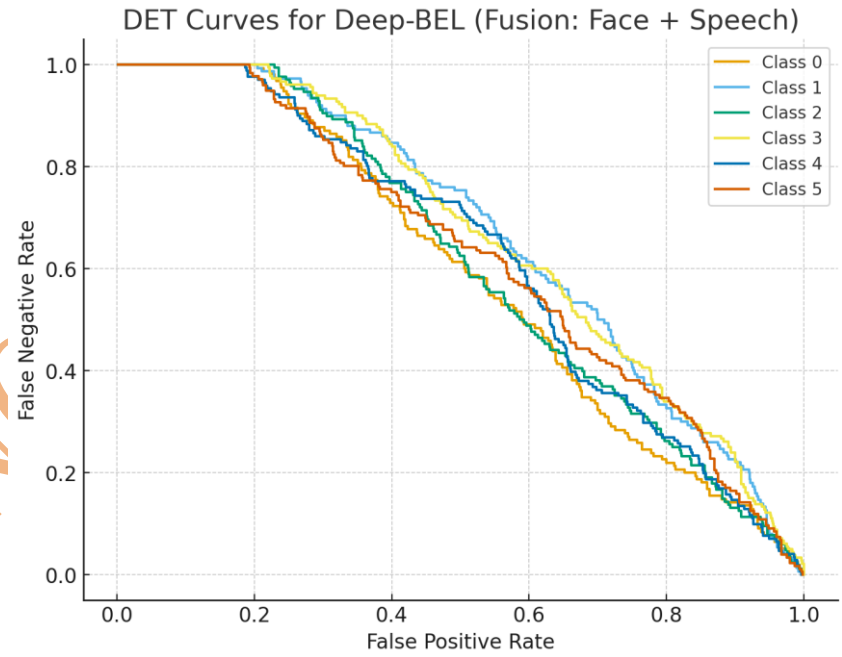
مرجع	سال	مدالیته‌ها	روش/مدل اصلی	دیتاست	دقت(%)
[30]	۲۰۲۳	چهره + گفتار	CNN + Attention Fusion	IEMOCAP, CMU-MOSEI	~۹۰
[31]	۲۰۲۴	چهره + گفتار	LSTM + Fusion	eNTERFACE’05, EmoDB	~۹۲
[32]	۲۰۲۴	چهره + گفتار	Transformer + Cross-Attention	CMU-MOSEI	۹۳
[33]	۲۰۲۴	گفتار + چندمدالیته	wav2vec 2.0 + Fusion	IEMOCAP	۹۴
[34]	۲۰۲۵	چهره + گفتار + بدن	Deep Fusion + Shared Module	AFEW, RAVDESS	۹۵
Deep-BEL	۲۰۲۵	چهره مرئی + گفتارIR +	CNN + BEL الهام از لیمبیک	Enterface’05	۹۶.۲۰

شکل ۳ عملکرد مدل را در جداسازی کلاس‌های هیجانی نشان می‌دهد. هرچه مساحت زیر منحنی (AUC) به عدد ۱ نزدیک‌تر باشد، مدل قدرت بیشتری در تفکیک کلاس‌ها دارد. همان‌طور که مشاهده می‌شود، مدل Deep-BEL در اکثر حالات هیجانی عملکردی قابل‌قبول داشته و منحنی‌ها به سمت گوشه بالا-چپ متمایل هستند. این موضوع بیانگر نرخ بالای تشخیص صحیح و نرخ پایین خطای مدل است.

UNCORRECTED PROOF



شکل ۳- منحنی‌های ROC برای شش حالت هیجانی (عصبانیت، تنفر، ترس، خوشحالی، ناراحتی و تعجب) در مدل Deep-BEL مبتنی بر همجوشی چهره و گفتار



شکل ۴- نمودار DET برای شش حالت هیجانی در مدل Deep-BEL مبتنی بر همجوشی چهره و گفتار

تصمیم‌گیری تطبیقی را به‌همراه دارد. به همین دلیل، این مدل می‌تواند نسبت به رویکردهای سنتی و تک‌مدالیته، عملکرد بهتری در بازشناسی هیجان داشته باشد.

۵- نتیجه‌گیری و پیشنهادات آتی

در این پژوهش، روشی ترکیبی با عنوان Deep-BEL برای بازشناسی هیجان مبتنی بر همجوشی چندمدالیته ارائه شد. در این مدل، سه ورودی شامل تصاویر مرئی چهره، تصاویر مادون قرمز و سیگنال‌های گفتار پس از پیش‌پردازش به شبکه‌های کانولوشنی ارسال شدند تا بازنمایی‌های عمیق ویژگی‌ها استخراج گردد. سپس ویژگی‌های حاصل در سطح ویژگی با استفاده از روش همبستگی کانونی و جمع‌برداری با یکدیگر ادغام شده و در ادامه به بخش‌های آمیگدال و اوربیتوفرانتال مدل BEL منتقل شدند تا طبقه‌بندی نهایی هیجان‌ها صورت گیرد. نتایج آزمایش‌ها بر روی پایگاه داده Enterface’05 نشان داد که مدل پیشنهادی با دستیابی به دقت ۹۴.۲۰ درصد نسبت به سایر روش‌های همجوشی موجود عملکرد برتری داشته است و این امر بیانگر کارایی ترکیب شبکه‌های کانولوشنی با معماری BEL و به‌کارگیری همزمان سه مدالیته در بهبود بازشناسی هیجان است.

با وجود این دستاوردها، مدل پیشنهادی همچنان با چالش‌هایی روبه‌رو است. از جمله می‌توان به محدودیت تعمیم‌پذیری ناشی از ارزیابی تنها بر روی یک پایگاه داده، پیچیدگی بالای محاسباتی و نیاز به منابع پردازشی قوی، استفاده از تصاویر مادون قرمز تولیدشده به جای داده‌های واقعی، و دشواری همزمان‌سازی کامل میان مدالیته‌های گفتار و چهره در شرایط طبیعی اشاره کرد. علاوه بر این، مانند سایر روش‌های مبتنی بر یادگیری عمیق، تفسیرپذیری بخش CNN نیز محدود بوده و تحلیل دقیق تصمیم‌گیری مدل را دشوار می‌سازد.

بر همین اساس، در تحقیقات آینده می‌توان مسیرهایی همچون ارزیابی مدل بر روی پایگاه‌های داده متنوع‌تر و در محیط‌های واقعی، استفاده از داده‌های مادون قرمز واقعی به‌منظور ارتقای قابلیت تعمیم، بهینه‌سازی ساختار محاسباتی برای کاهش پیچیدگی و هزینه پردازش، طراحی الگوریتم‌های همجوشی تطبیقی برای مدیریت عدم همزمانی مدالیته‌ها و بهره‌گیری از روش‌های نوین توضیح‌پذیری در یادگیری عمیق به‌منظور افزایش شفافیت فرایند تصمیم‌گیری مدل را دنبال کرد.

شکل ۴، نمودار DET رابطه بین نرخ خطای مثبت کاذب و منفی کاذب را روی مدل پیشنهادی نمایش می‌دهد. نقاطی که منحنی‌ها به محور قطری نزدیک می‌شوند، بیانگر شرایطی هستند که نرخ خطاهای مدل در تعادل قرار می‌گیرند. این نقطه به‌عنوان نرخ خطای برابر شناخته می‌شود. بر اساس نتایج، مقدار نرخ خطای برابر، برای حالات مختلف در بازه ٪۵۵ تا ٪۶۱ قرار دارد که نشان‌دهنده کارایی قابل‌قبول مدل در تفکیک هیجانات است.

انتخاب مدل Deep-BEL در این پژوهش بر پایه ملاحظات فنی متعددی صورت گرفته است. نخست، شبکه‌های عصبی کانولوشنی (CNN) به دلیل توانایی بالای خود در استخراج بازنمایی‌های سلسله‌مراتبی و محلی از داده‌های دیداری و شنیداری، یکی از پرکاربردترین معماری‌ها در حوزه تشخیص الگو به‌شمار می‌آیند. استفاده از CNN در بخش تالاموس مدل پیشنهادی باعث می‌شود تا ویژگی‌های سطح پایین و میانی تصاویر مرئی، مادون قرمز و مل‌اسپکتروگرام‌های گفتار با دقت بیشتری استخراج گردند، ضمن اینکه لایه‌های Pooling با کاهش ابعاد و پارامترها، پیچیدگی محاسباتی را کاهش داده و امکان پردازش کارآمدتر داده‌های بزرگ را فراهم می‌سازند.

از سوی دیگر، انتخاب مدل یادگیری هیجانی مغز(BEL) به دلیل شبیه‌سازی ساختارهای کلیدی سیستم لیمبیک مانند آمیگدال و اوربیتوفرانتال بوده است. این مدل امکان پردازش تقویتی مبتنی بر پاداش و جریمه را فراهم می‌آورد که شباهت زیادی به یادگیری هیجانی انسان دارد و در مسائل تصمیم‌گیری پیچیده کارایی بالایی نشان داده است. افزودن CNN به بخش‌های تالاموس و قشر حسی مدل BEL، باعث ارتقای توان استخراج ویژگی و افزایش دقت طبقه‌بندی می‌شود.

همچنین استفاده از همجوشی در سطح ویژگی با کمک روش CCA یک انتخاب فنی هدفمند است، زیرا این روش قادر است همبستگی بین بردارهای ویژگی چندمدالیته را استخراج کرده و ترکیب بهینه‌تری از اطلاعات شنیداری و دیداری ارائه دهد. این امر به‌ویژه در شرایطی که مدالیته‌ها دارای نویز یا عدم همزمانی جزئی باشند، به بهبود کارایی طبقه‌بندی کمک می‌کند.

در نهایت، معماری پیشنهادی Deep-BEL ترکیبی از مزیت‌های یادگیری عمیق در استخراج بازنمایی‌های غنی و مزیت‌های مدل‌های بیولوژیک‌الهام‌گرفته در همجوشی و

UNCORRECTED PROOF

مراجع

[1] A. Mehrabian, “Communication without words,” Psychology Today, vol. 2, pp. 53-56, 1968.
[2] N. Sebe, I. Cohen, T. Gevers, T. S. Huang, “Multimodal approaches for emotion recognition: a survey,” Proceedings of the SPIE, Vol 5670, pp. 56-67, 2004.
[3] T. Wark and S. Sridharan, “Adaptive fusion of speech and lip information for robust speaker identification,” Digital Signal Processing, 2001.
[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, pp. 39-58, 2009.
[5] Shivappa, S.T.; Trivedi, M.M.; Rao, B.D., “Audiovisual Information Fusion in Human–Computer Interfaces and Intelligent Environments: A Survey,” Proceedings of the IEEE, vol.98, no.10, pp.1692, 1715, Oct. 2010.
[6] Sidney K. D'Mello, Jacqueline M. Kory, “A Review and Meta-Analysis of Multimodal Affect Detection Systems,” ACM Comput. Surv, vol. 47(3), pp.1-36, 2015.
[7] P. Ekman, W. V. Friesen, and J. C. Hager, Facial Action Coding System (FACS), The Manual: A Human Face, 2002.
[8] Y. Chang, C. Hu, R. Feris, and M. Turk, “Manifold Based Analysis of Facial Expression,” J. Image and Vision Computing, vol. 24, no. 6, pp. 605-614, 2006.
[9] M. Pantic and I. Patras, “Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments Form Face Profile Image Sequences,” IEEE Trans. Systems, Man, and Cybernetics Part B, vol. 36, no. 2, pp. 433-449, 2006.
[10] G. Guo and C. R. Dyer, “Learning from examples in the small sample case: Face expression recognition,” IEEE Trans. Systems, Man and Cybernetics–Part B, vol. 35, no. 3, pp. 477-488, 2005.
[11] K. Anderson and P. W. McOwan, “A real-time automated system for recognition of human facial expressions,” IEEE Trans. Systems, Man and Cybernetics–Part B, vol. 36, no. 1, pp. 96-105, 2006.
[12] J. Whitehill and C. W. Omlin, “Haar features for FACS AU recognition,” in Proc. IEEE Int. Conf. Autom. Face and Gesture Recog., pp. 217-222, 2006.
[13] M. Pantic and I. Patras, “Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments Form Face Profile Image Sequences,” IEEE Trans. Systems, Man, and Cybernetics Part B, vol. 36, no. 2, pp. 433-449, 2006.
[14] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 915–928, 2007.
[15] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, “Emotion recognition using phog and lpq features,” in 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, 2011.
[16] N. Sebe, I. Cohen, T. Gevers, T. Huang, “Multimodal Approaches for Emotion Recognition: A Survey”, in: Proceedings of SPIE, 2005.
[17] M. Ayadi, M. S. Kamel, F. Karray, "Speech Emotion Recognition Using Gaussian Mixture Vector Autoregressive Models”. International Conference on Acoustics, Speech, and Signal Processing, 2007.
[18] F. Dellaert, T. Polzin, A. Waibel, ” Recognizing emotion in speech. *International conference on spoken language processing”,* 1970-1973, 1996.
[19] A. Harimi, A. Shahzadi, A. Ahmadyfard, K.Yaghmaie, “Classification of emotional speech spectral pattern features”. Journal of AI and Data Mining, *2*(1), 53-61, 2014.
[20] G.Rázuri, J. Sundgren, D. Rahmani, R. Larsson, “Speech emotion recognition in emotional feedback for human - robot interaction”, International Journal of Advanced Research in Artificial Intelligence, 4, 20-27, 2015.
[21] J. Winter, Y. Xu, W.C. Lee, ”Energy efficient processing of k-nearest neighbor queries location aware sensor networks”. The Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services, San Diego,CA, 281-292, 2005.
[22] M. E. Ayadia, M. S. Kamelb, F. Karraybc, “Survey on speech emotion recognition: features, classification schemes and databases”, Pattern Recognition, 572- 587, 2011.
[23] M. Sheikhan, M. Bejani, D. Gharavian, “Modular Neural-SVM Scheme for Speech Emotion Recognition Using ANOVA Feature for Method”, Neural Computing and Applications, *23*, 2012.
[24] S. U. Haq, P. J. B. Jackson, J. Edge, “Audio-visual feature selection and reduction for emotion classification”. International Conference on Auditory-Visual Speech Processing, 2008.
[25] D. Ververidis, C. Kotropoulos, “Emotional speech recognition: resources, features and methods”, Elsevier Speech communication, *48*(9), 1162- 1181, 2006.
[26] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou, "End-to-end speech emotion recognition using a deep convolutional recurrent network", In Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal

ACCEPTED PROOF

Processing (ICASSP), Shanghai, China, pp 5200–5204, 2016.
[27] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks", In Proceedings of the Twenty-Sixth Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, pp 1097–1105, 2012.
[28] A. M. Badshah, J. Ahmad, N. Rahim, S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network", In Proceedings of the International Conference on Platform Technology and Service (PlatCon), Busan, Korea, pp 1–5, 2017.
[29] M. Mansoorizadeh, Nasrollah Moghaddam Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," Multimedia Tools Appl. 49(2): 277-297, 2010.
[30] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features," *Sensors*, vol. 23, no. 12, p. 5475, 2023, doi: 10.3390/s23125475.
[31] J. Salas-Cáceres, J. Lorenzo-Navarro, D. Freire-Obregón, and M. Castrillón-Santana, "Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics," *Multimedia Tools and Applications*, vol. 84, pp. 27327–27343, 2025, doi: 10.1007/s11042-024-20227-6.
[32] P. Waligora, M. H. Aslam, M. O. Zeeshan, S. Belharbi, A. L. Koerich, M. Pedersoli, S. Bacon, and E. Granger, "Joint Multimodal Transformer for Emotion Recognition in the Wild," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2024, pp. 4625–4635, doi: 10.1109/CVPRW63382.2024.00465.
[33] F. Li, J. Luo, and W. Xia, "WavFusion: Towards wav2vec 2.0 Multimodal Speech Emotion Recognition," *arXiv preprint* arXiv:2412.05558, 2024.
[34] S. Fan, J. Jing, and C. Wang, "Audio-Visual Learning for Multimodal Emotion Recognition," *Symmetry*, vol. 17, no. 3, p. 418, 2025, doi: 10.3390/sym17030418.
[35] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 2, no. 4, pp. 53–56, 1968.

UNCORRECTED PROOF