سیستم های فازی و کاربردها
سال هشتم، شماره دوم (پائیز و زمستان ۱۴۰۴)
صص. ۴۳ تا ۶۳
شاپای چاپی: ۲۷۱۷ـ۴۴۰۹
شاپای الکترونیکی: ۲۷۱۷ـ۳۹۸۴
DOI:10.22034/jfsa.2025.506704.1262

همجوشی فازی مدل‌های یادگیری ماشین برای تشخیص احساسات با استفاده از معماری انتزاع و همجوشی تصمیم

محمد صفائی، سمیه مغاری* و مریم جلالی
دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود، شاهرود، ایران

تاریخ دریافت: ۱۴۰۳/۱۱/۲۶ تاریخ پذیرش: ۱۴۰۴/۰۷/۰۹

نوع مقاله: علمی‌ـ پژوهشی

چکیده. تشخیص احساسات در محاسبات عاطفی نقش مهمی در بهبود تعامل انسان و ماشین، نظارت بر سلامت روان و ایجاد تجربیات کاربری شخصی‌سازی‌شده دارد. با توجه به اهمیت رو به رشد این حوزه، این مقاله به معرفی یک مدل یادگیری ماشین توسعه داده شده بر پایه معماری انتزاع و همجوشی تصمیم برای تشخیص احساسات می‌پردازد. این معماری در سه لایه انتزاع، محاسبه و همجوشی تعریف شده است و نوآوری مدل پیشنهادی پیرامون طراحی مؤلفه‌ها و آرایش این لایه‌ها است. در این راستا، برای لایه نخست، دو نوع انتزاع برای خلاصه‌سازی ویدئوها تعریف شده و سپس در لایه محاسبه، دو مدل پایه سبک شامل یک شبکه عصبی اتصال کامل و یک شبکه عصبی پیچشی با عمق کم طراحی و پیاده‌سازی شده‌اند. در لایه همجوشی، تصمیم مدل‌های پایه با استفاده از منطق فازی ترکیب شده تا مدل بتواند تصمیم نهایی را بگیرد. این مدل روی مجموعه داده‌های L-SVD برای کلاس‌بندی ویدئوهای تشخیص احساسات به دقت بیش از ۹۶ درصد دست یافته است. نتایج تجربی نشان می‌دهند که مدل پیشنهادی نه تنها در دقت تشخیص، بلکه از جنبه‌های بار پردازشی و حافظه مصرفی نیز به طور قابل توجهی از مدل‌های مدرن موجود بهتر است.

2010 *Mathematics Subject Classification.* 68T45; 26E50
* Corresponding author
E-mails: s.moghari@shahroodut.ac.ir, mohamad.safaei@shahroodut.ac.ir and m.jalali@shahroodut.ac.ir.
عبارات و کلمات کلیدی. محاسبات عاطفی، یادگیری ماشین، معماری انتزاع و همجوشی تصمیم، همجوشی فازی.

۱. مقدمه

احساسات، در تصمیم‌گیری، ادراک، یادگیری و بسیاری از مکانیسم‌های تفکر منطقی نقش اساسی ایفا می‌کنند [۲۰]. درک احساسات در حوزه سلامت روان نیز اهمیت ویژه دارد؛ زیرا به تشخیص و درمان به‌موقع مشکلات روانی کمک می‌کند و می‌تواند به افراد در مدیریت استرس و اضطراب یاری رساند. از سوی دیگر، برای آنکه رایانه‌ها و تجهیزات هوشمند توانایی برقراری ارتباط طبیعی با ما را باشند، باید بتوانند هیجانات را تشخیص داده، درک کرده و حتی تجربه و نمایش دهند. محاسبات عاطفی[1]، که به بررسی بین‌رشته‌ای احساسات و محاسبات می‌پردازد، اهمیت تشخیص عواطف انسانی در کاربردهای گوناگون مانند نظارت بر سلامت روان، تجربیات شخصی‌سازی‌شده کاربر و همچنین تعامل انسان و رایانه را مورد توجه قرار داده است [۱۵]. یکی از کاربردهای مدرن در این زمینه، تجهیزات پوشیدنی هوشمند است که فرصتی برای ادغام محاسبات عاطفی در زندگی روزمره انسان‌ها را به وجود آورده است [۱۶]. این ابزارها، پایش مداوم و غیرمزاحم حالات احساسی را فراهم می‌آورند.

با توجه به محدودیت منابع در کاربردهایی مانند دستگاه‌های پوشیدنی، بهینه‌سازی محاسبات در راستای تسریع پردازش اطلاعات [۱، ۷] منجر به بهبود عملکرد کلی و پاسخگویی در زمان واقعی برنامه‌هایی مانند نظارت بر سلامت و تشخیص فعالیت و احساسات می‌شود. در این راستا، یکی از جنبه‌های کلیدی بهینه‌سازی محاسباتی و افزایش کارایی، کاهش بار کاری داده‌ها است. این بُعد، شامل تکنیک‌هایی است که بر کاهش حجم و پیچیدگی داده‌هایی که نیاز به پردازش دارند تمرکز دارد، بنابراین بهره‌وری محاسباتی را افزایش می‌دهد.

معماری انتزاع و همجوشی تصمیم (ADFA[2])، رویکردی نوین در توسعه مدل‌های ترکیبی یادگیری ماشین برای غلبه بر چالش‌هایی مانند بار محاسباتی بالا و محدودیت فضای ذخیره سازی ارائه داده است [۶]. در لایه نخست معماری ADFA، یعنی لایه انتزاع، مجموعه‌ای از روال‌های خلاصه‌سازی داده‌ها تعریف می‌شوند که داده‌ها را براساس دیدگاه‌های خاصی مورد پردازش قرار می‌دهند تا حجم قابل توجهی از آن را کاهش دهند. لایه دوم، محاسبه نام دارد و از تعدادی مدل محاسباتی سبک‌وزن تشکیل می‌شود که هر یک به‌صورت مستقل یک انتزاع داده را پردازش و تصمیم‌گیری می‌کنند. در لایه سوم که همجوشی نام دارد، خروجی‌های مدل‌های لایه دوم تجمیع می‌شوند تا خروجی نهایی را شکل دهند. فرآیند همجوشی با ترکیب مدل‌های به دست آمده از لایه محاسبه، تعادلی مابین دقت، کارآیی و میزان مصرف منابع را برقرار می‌کند.

---

[1] Affective Computing

[2] Abstraction and Decision Fusion Architecture

در این پژوهش، یک مدل یادگیری ماشین جدید بر پایه معماری انتزاع و همجوشی تصمیم برای تشخیص احساسات از روی حالت چهره در داده‌های ویدئویی ارائه شده است. مدل پیشنهادی، از دو روال انتزاع داده، دو مدل محاسباتی سبک‌وزن و یک سیستم استنتاج فازی استفاده می‌کند. این مدل، با تمرکز بر بهینه‌سازی عملکرد و استفاده هوشمندانه از منابع توسعه یافته و می‌تواند مشکلاتی که در زمینه پیچیدگی محاسباتی و محدودیت فضای ذخیره سازی بود را با حفظ دقت تشخیص بالا در زمینه تشخیص احساسات به‌طور چشم‌گیری کاهش دهد.

در ادامه، در بخش۲ پژوهش‌های مرتبط در زمینه تشخیص احساسات مورد بررسی قرار گرفته است. بخش ۳ به معرفی مدل پیشنهادی پرداخته و جزئیات ساختار آن در قالب لایه‌های مختلف معماری ADFA را ارائه می‌دهد. در بخش ۴، نتایج حاصل از آزمایش مدل پیشنهادی ارائه شده و نتایج به دست آمده با مدل‌های مدرن تشخیص احساسات مقایسه شده است. در پایان، بخش ۵ به جمع‌بندی و نتیجه‌گیری نتایج و دستاوردهای این پژوهش می‌پردازد.

۲. پیشینه پژوهش

تشخیص احساسات امکان درک بهتر نیازها و در نتیجه پاسخ بهتر به آنها را فراهم می‌کند. برای نمونه، می‌تواند در بهبود خدمات مشتریان و تجربه کاربری در محصولات و خدمات آنلاین نقش کلیدی ایفا کند [۱۸]. در این راستا، تکنیک‌های مختلفی به کار گرفته شده‌اند تا بتوانند احساسات را از متون، تصاویر، ویدیوها و سایر داده‌ها استخراج کنند. جدول ۱ دسته‌بندی کلی انواع مدل‌های یادگیری ماشین که در تشخیص احساسات به کار گرفته شده‌اند را نمایش می‌دهد. نمونه‌ها در قالب چهار دسته مدل‌های ساده، مدل‌های ترکیبی، شبکه‌های عصبی و مبدل‌ها گروه‌بندی شده‌اند و به نقاط قوت و ضعف آنها اشاره شده است.

جدول ۱: دسته‌بندی مدل‌های یادگیری به کار رفته در ماشین تشخیص احساسات.

دسته	نمونه	نقطه قوت	نقطه ضعف
مدل‌های ساده	ماشین بردار پشتیبان	عملکرد خوب در داده‌های کوچک	رشد اندازه و محاسبات با داده‌های آموزش
	درخت تصمیم	عملکرد خوب در داده‌های کوچک	دقت پایین در دسته‌بندی داده‌های پیچیده
مدل‌های ترکیبی	شبکه مولد خصمانه	تولید داده‌های مصنوعی با تنوع بالا	پیچیدگی آموزش و مصرف زیاد منابع
	یادگیری گروهی	بهبود دقت نسبت به مدل‌های ساده	نیاز به منابع محاسباتی و حافظه زیاد
	تقویت سازگار	بهبود دقت نسبت به مدل‌های ساده	کم بودن دقت در داده‌های پیچیده
	جنگل تصادفی	بهبود دقت نسبت به مدل‌های ساده	زمان اجرای طولانی و مصرف حافظه
شبکه‌های عصبی	شبکه‌های بازگشتی	درک داده‌های ترتیبی و سری زمانی	پیچیدگی معماری
	شبکه‌های پیچشی	دقت بالا، استخراج خودکار ویژگی‌ها	پیچیدگی محاسباتی و حافظه
	شبکه‌های ژرف	دقت بالا، استخراج خودکار ویژگی‌ها	پیچیدگی محاسباتی و حافظه
مبدل‌ها	مبدل بصری	پردازش موازی داده‌های ترتیبی	پیچیدگی محاسباتی و حافظه

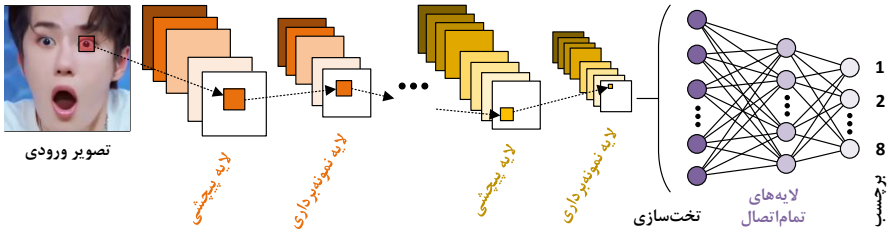
مدل‌های ساده مانند ماشین بردار پشتیبان (SVM[1]) و درخت تصمیم[2]، علاوه بر ضعف در دسته‌بندی داده‌های پیچیده، با بزرگ شدن اندازه مجموعه داده‌های آموزش، مدل نهایی نیز رشد می‌کند و از نظر مصرف حافظه و منابع پردازشی سنگین می‌شود [۲]. مدل‌های ترکیبی مانند شبکه مولد خصمانه[3]، یادگیری گروهی[4]، تقویت سازگار[5] و جنگل تصادفی[6]، نسبت به مدل‌های ساده دارای قابلیت بیشتر برای درک داده‌های پیچیده مانند تصاویر و ویدئوها دارند و به همین دلیل دقت آنها بالاتر است. البته این مدل‌ها بار پردازشی بیشتر و مصرف منابع زیادتر را به سیستم تحمیل می‌کنند و برای پردازش‌های منابع محور یا زمان واقعی مناسب نیستند [۵، ۱۲]. از سوی دیگر،شبکه‌های عصبی بازگشتی (RNN[7])، شبکه‌های عصبی پیچشی (CNN[8]) و شبکه‌های ژرف[9] توانایی استخراج خودکار ویژگی‌ها و دسته‌بندی داده‌های پیچیده با دقت خوب را دارند اما از نظر مصرف منابع، مدل‌های سنگین به شمار می‌روند. مبدل‌های بصری[10] نیز نقطه ضعف شبکه‌های عصبی بازگشتی در پردازش ترتیبی داده‌ها را برطرف کرده و می‌توانند داده‌های ترتیبی و سری‌های زمانی را به صورت موازی پردازش کنند. البته این مدل‌ها نیز به منابع پردازشی و حافظه زیاد نیاز دارند [۲۳].

بیشترین مدل‌های یادگیری ماشین مورد استفاده برای تشخیص احساسات مبتنی برشبکه‌های عصبی پیچشی و شبکه‌های عصبی ژرف [11] هستند. این شبکه‌ها با استفاده از فیلترهای لایه پیچشی توانایی بالایی در استخراج خودکار ویژگی‌ها از داده‌های خام دارند [۲، ۱۹]. شبکه‌های عصبی ژرف مانند شبکه همجوشی حافظه [12]، شبکه همجوشی تنسوری[13] و شبکه ترجمه چرخه‌ای چندوجهی[14] به دلیل توانایی بالای این شبکه‌ها در استخراج ویژگی‌های مکانی و زمانی کاربرد گسترده‌ای دارند [۲۲]. شبکه عصبی ژرف ResNet-18 دارای ۱۸ لایه شامل لایه‌های پیچشی، اتصالات اضافی و لایه‌های کاملا متصل است، که به علت استفاده از اتصالات اضافی و کاهش ناپایداری گرادیان، دقت بالایی در کلاس‌بندی تصاویر دارد. این شبکه برای تشخیص احساسات

[1] Support Vector Machine
[2] Decision Tree
[3] Generative Adversarial Network
[4] Ensemble Learning
[5] Adaptive Boosting
[6] Random Forest
[7] Recurrent Neural Networks
[8] Convolutional Neural Networks
[9] Deep Networks
[10] Vision Transformers
[11] Deep Neural Networks
[12] Memory Fusion Network
[13] Tensor Fusion Network
[14] Multimodal Cycle-Consistent Generative Adversarial Networks

از مجموعه داده‌های تصویری به کار گرفته شده است [۹]. شبکه پیچشی نقطه‌ای[1] با اعمال فیلترهای یک‌دریک روی هر کانال ورودی به‌طور جداگانه، باعث کاهش تعداد پارامترها، کاهش بار محاسباتی و افزایش سرعت آموزش می‌شود؛ هرچند به‌تنهایی قادر به استخراج ویژگی‌های پیچیده نیست [۳]. شبکه تعبیر احساس مبتنی بر بینایی [۲۳] هم از تکنیک‌های یادگیری ژرف استفاده می‌کند تا ویژگی‌های بصری و زمانی را از شبکه‌های عصبی بازگشتی استخراج کند.

شکل ۱ ساختار کلی شبکه‌های عصبی پیچشی را نمایش می‌دهد. این مدل‌ها دارای لایه‌های پیچشی متعدد هستند که سبب می‌شود ویژگی‌های پیچیده مانند لبه‌ها، زوایا و اشکال استخراج شوند. البته، این شبکه‌ها نیازمند مجموعه داده‌های بزرگ هستند تا بتوانند نتایج دقیق و قابل اعتماد تولید کنند [۳]. از سوی دیگر، با اینکه پس از هر لایه پیچشی، یک لایه نمونه‌برداری به کاهش بار داده و حجم پردازش‌های بعدی کمک می‌کند، اما استفاده از تعداد قابل توجهی فیلتر در لایه‌های پیچشی متعدد و تولید نقشه‌های ویژگی به صورت تنسورهای بزرگ، آنها را به مدل‌های پردازشی سنگین تبدیل می‌کند.



شکل ۱: ساختار کلی شبکه‌های عصبی پیچشی.

شبکه‌های عصبی بازگشتی برای مدل‌سازی دنباله‌های سری زمانی طراحی شده‌اند و می‌توانند روابط بین فریم‌های ویدئو را مدل‌سازی کنند؛ اما در ویدئوهای طولانی و دنباله‌های بلندمدت دچار مشکل گریز از گرادیان می‌شوند [۱۲]. شبکه‌های عصبی بازگشتی حافظه کوتاه‌مدت طولانی (LSTM[2])، مشکل گریز از گرادیان را تا حدودی حل می‌کنند و توانایی حفظ و یادگیری اطلاعات بلندمدت را دارند. البته، این شبکه‌ها به‌علت داشتن ساختار پیچیده، بار محاسباتی بالایی را به سیستم تحمیل می‌کنند و بر خلاف مبدل‌ها، قابلیت پردازش موازی اطلاعات دنباله‌ای را ندارند که موجب کارایی پایین‌تر آنها می‌شود [۱۴].

[1] Pointwise Convolution
[2] Long Short-Term Memory

در مدل ترکیبی شبکه عصبی پیچشی و مبدل، ابتدا شبکه عصبی ویژگی‌های مکانی را استخراج می‌کند. سپس، ویژگی‌های استخراج شده به مبدل داده می‌شود تا تحلیل زمانی انجام شود. [۱۱]. مدل‌های ترکیبی شبکه عصبی با ترکیب چندین معماری یا روش مختلف سعی در بهبود عملکرد و دقت پیش‌بینی دارند. مدل پیچشی مکانی‌ـ‌زمانی[1]، شامل لایه‌های پیچشی مکانی، لایه‌های پیچشی زمانی و لایه‌های ترکیبی هستند. لایه‌های پیچشی مکانی، اطلاعات مکانی را از هر فریم استخراج می‌کند. لایه‌های پیچشی زمانی، تغییرات در طول زمان را تحلیل می‌کند. این مدل به دلیل پردازش هم‌زمان اطلاعات مکانی و زمانی، از یک سو قدرت بالاتری نسبت به مدل‌های غیرترکیبی دارد و از طرفی به‌علت داشتن لایه های بیشتر نیازمند داده‌های زیاد برای آموزش است و همچنین مصرف منابع بالایی دارد [۲۴]. شبکه پیچشی بازگشتی بلندمدت[2]، ترکیبی از شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی است. در این مدل، ویژگی‌های مکانی در فریم‌های ویدئویی با شبکه‌های عصبی پیچشی استخراج می‌شوند. سپس ویژگی‌های استخراج شده از هر فریم به شبکه‌های بازگشتی داده می‌شود تا اطلاعات زمانی و ترتیبی بین فریم‌ها بررسی شوند. قدرت ترکیبی این دو شبکه باعث می‌شود دقت مدل بهبود یابد؛ اما ترکیب دو نوع شبکه عصبی، پیچیدگی محاسباتی را افزایش می‌دهد [۱۳].

به کارگیری مدل‌های یادگیری ماشین مانند رگرسیون لجستیک[3] و جنگل تصادفی برای کلاس بندی داده‌های تشخیص احساسات [۵، ۴] کمتر مورد استقبال قرار گرفته است؛ زیرا رگرسیون لجستیک قابلیت بالایی برای مدل‌سازی روابط غیرخطی پیچیده را ندارد و همچنین تفسیر ضرایب رگرسیون لجستیک، به‌ویژه در مسائل چندبعدی، بسیار دشوار است [۱۷، ۲۰]. همچنین، در زمان آموزش مدل ماشین بردار پشتیبان، پارامترهای زیادی تولید می‌شوند که باعث می‌شود نیازمند عملیات پردازشی شود. مدل‌هایی مانند ماشین بردار پشتیبان که به تعداد داده‌های ورودی آموزش حساس هستند، چند مسئله اساسی را به وجود می‌آورند. نخست اینکه با افزایش تعداد داده‌های آموزشی، به ویژه در مسائل با ابعاد بالا، از نظر محاسباتی کُند و پیچیده می‌شوند. بعلاوه، اندازه مدل به تعداد بردارهای پشتیبان وابسته است که با افزایش تعداد داده‌های آموزشی، منجر به افزایش حافظه مورد نیاز و پیچیدگی مدل می‌شود [۲، ۸].

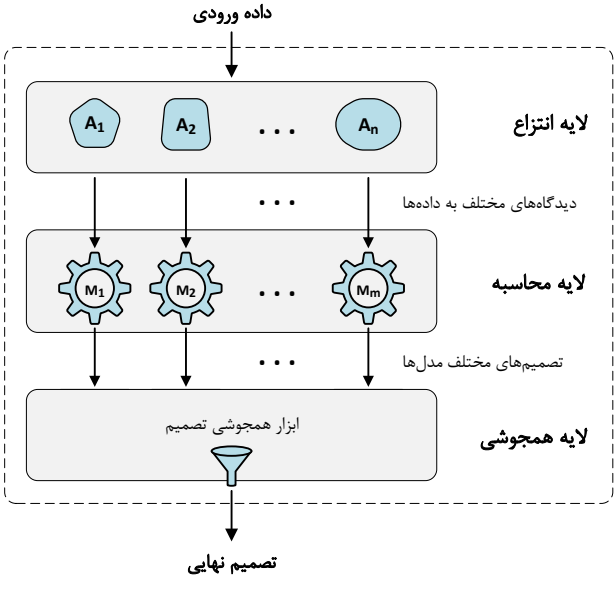
در این پژوهش، به دنبال توسعه مدل یادگیری ماشینی برای تشخیص احساسات هستیم که علاوه بر دقت بالا، بار پردازشی و مصرف منابع آنها کم باشد و همچنین برای آموزش، نیاز به مجموعه داده بزرگ نداشته باشد.

---

[1] Spatiotemporal Convolutional Neural Network
[2] Long-term Recurrent Convolutional Network
[3] Logistic Regression

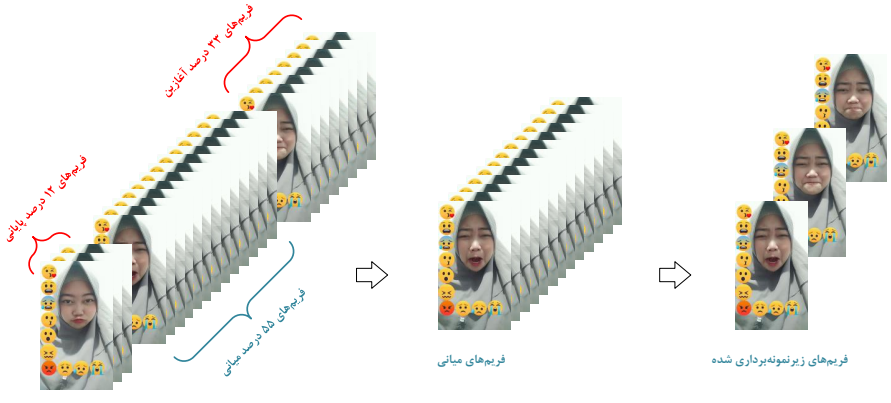
۳. رویکرد پیشنهادی

در این پژوهش، توسعه مدل‌های یادگیری ماشین برپایه معماری ADFA در سه لایه مفهومی انتزاع، محاسبه و همجوشی انجام می‌شود [۶]. ساختار کلی این معماری در شکل ۲ نمایش داده شده است. در لایه انتزاع، داده‌ها براساس دیدگاه‌های مختلف خلاصه‌سازی می‌شوند تا حجم آنها و در نتیجه بار کاری پردازش کاسته شود. بنابراین، در این لایه چند نسخه متفاوت و کم‌حجم از داده ورودی تولید می‌شود. خروجی هر روال انتزاع، توسط یک یا چند مدل محاسباتی در لایه دوم پردازش می‌شود. به عبارت دیگر، هر نسخه خلاصه شده داده در لایه اول، به عنوان ورودی یک یا چند مدل محاسباتی در لایه دوم مورد استفاده قرار می‌گیرد. مدل‌ها در این لایه سبک‌وزن و مستقل از یکدیگر هستند و خروجی آنها در لایه سوم توسط یک ابزار یا روش همجوشی، با یکدیگر آمیخته می‌شود تا خروجی مدل ADFA (تصمیم نهایی) تولید شود. هدف از همجوشی تصمیم مدل‌ها، افزایش دقت کلی با ترکیب تشخیص‌هایی است که در نتیجه دیدگاه‌های متفاوت به داده‌ها گرفته شده‌اند. این معماری، از یک سو با کاهش داده‌ها، امکان استفاده از مدل‌های سبک‌وزن و در نتیجه کاهش میزان پردازش و مصرف منابع را فراهم می‌آورد، و از سوی دیگر، با همجوشی تصمیم مدل‌ها، دقت از دست رفته را جبران می‌کند.



شکل ۲: ساختار کلی معماری ADFA [۱، ۶] مبنای توسعه مدل پیشنهادی.

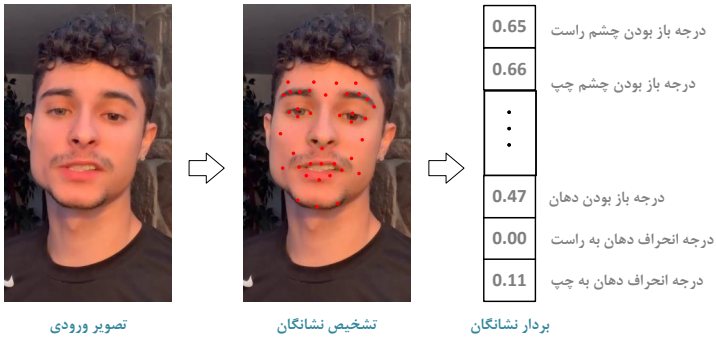
۱.۳. لایه انتزاع. در لایه نخست مدل پیشنهادی، دو انتزاع به نام‌های نشانگان[1] و ناحیه موردعلاقه[2] را توسعه داده‌ایم. البته هر دو انتزاع از یک روال مشترک زیرنمونه‌برداری[3] استفاده می‌کنند که در شکل ۳ نمایش داده شده است. در این روال، ابتدا ٪۳۳ از فریم‌های آغازین و ٪۱۲ از فریم‌های پایانی ویدئو نادیده گرفته می‌شود و سپس از ٪۵۵ باقی‌مانده، تعدادی فریم با فاصل یکنواخت انتخاب می‌شود. دلیل نادیده گرفتن فریم‌های اغازین و پایانی، خنثی بودن احساس در آنها است؛ زیرا در داده‌های مورد استفاده، در فریم‌های آغازین فرد به احساس مورد نظر نرسیده و در فریم‌های پایانی هم احساس را از دست داده و به حالت خنثی نزدیک شده است. قابل ذکر است که یافتن زمان شروع و پایان احساس، در این پژوهش مورد بررسی قرار نگرفته و درصدهای بیان شده صرفاً براساس تحلیل ویدئوهای دیتاست L-SVD [۲۱] است.



شکل ۳: روند زیرنمونه‌برداری فریم‌ها.

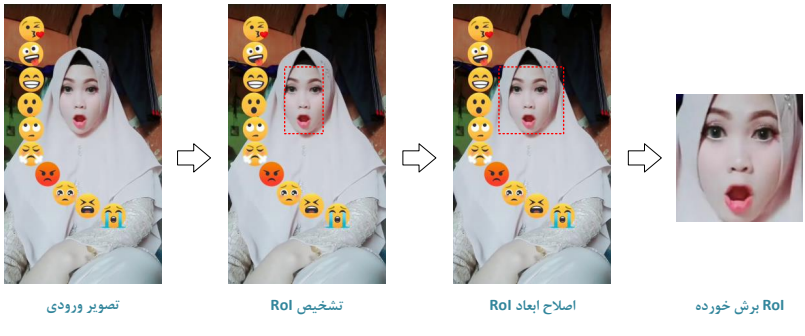
شکل ۴ مراحل انتزاع نشانگان روی یک فریم زیرنمونه‌برداری شده را نشان می‌دهد. در طی این مراحل، ابتدا نقاط نشانگان چهره مانند نقاط روی پیشانی، ابرو، دور چشم‌ها، گونه‌ها، بینی، دور لب‌ها، خطوط لبخند و اخم و چانه استخراج می‌شود. در ادامه، بردار نشانگان شامل ویژگی‌هایی مانند موقعیت نقاط مهم چهره و شدت حالت‌های چهره محاسبه می‌شود. درواقع، هر درایه از این بردار یک درجه فازی برای شدت باز بودن هر یک از چشم‌ها، انحراف هر کدام از ابروها، خمیدگی میانه ابروها، باز بودن دهان، انحراف دهان به چپ یا راست، خمیدگی دهان به بالا یا پایین، کشیدگی چانه و زاویه فک‌ها است. این بردار شامل اطلاعات مهمی پیرامون ویژگی‌های احساسی چهره است که به عنوان یک انتزاع از تصویر درنظر گرفته می‌شود.

[1] Landmark
[2] Region of Interest
[3] Subsampling



شکل ۴: انتزاع نشانگان روی یک فریم از ویدئو.

شکل ۵ مراحل انتزاع ناحیه موردعلاقه (RoI) روی یک فریم زیرنمونه‌برداری شده را نشان می‌دهد. هدف از این عملیات، حذف بخش‌های اضافی تصویر و کاهش حجم داده ورودی است. این کار، علاوه بر کاهش پردازش و مصرف منابع، استخراج ویژگی‌های مهم را برای مدل راحت‌تر می‌کند و در نتیجه دقت مدل محاسباتی که با این داده‌ها آموزش می‌بینند، افزایش می‌یابد. از آنجا که چهره حاوی اطلاعات مهم برای تحلیل حالت‌های احساسی است؛ ابتدا با یک مدل ساده ناحیه سر تشخیص داده می‌شود و سپس ابعاد این ناحیه اصلاح می‌شود. برای اصلاح ابعاد، عرض مستطیل تشخیص ناحیه از دو طرف اضافه می‌شود تا به مربع تبدیل شود. در ادامه، این ناحیه برش خورده و سپس به ابعاد ۱۰۰ × ۱۰۰ پیکسل تغییر اندازه داده می‌شود. این تصویر برش خورده با ابعاد استاندارد، خروجی انتزاع است که روی هر فریم از تصویر به صورت مجزا انجام می‌شود. البته در صورتی که فاصله فریم‌ها از یکدیگر زیاد نباشد، ناحیه برش را می‌توان پس از محاسبه در فریم نخست، در فریم‌های بعدی استفاده کرد.



شکل ۵: انتزاع ناحیه موردعلاقه روی یک فریم از ویدئو.

۲.۳. لایه محاسبه. در این لایه از دو مدل محاسباتی شبکه عصبی اتصال کامل (FCN[1]) و عصبی پیچشی با عمق کم[2] استفاده نموده‌ایم که به ترتیب از انتزاع نشانگان و انتزاع ناحیه موردعلاقه به عنوان داده ورودی آنها استفاده می‌شود. جدول ۲ مشخصات مدل FCN پیشنهادی را نمایش می‌دهد. این مدل دارای شش لایه اتصال کامل است که پنج لایه نخست از تابع فعال‌ساز ReLU و لایه آخر از تابع فعال‌ساز Softmax استفاده می‌کنند. این مدل از نظر پردازشی و مصرف حافظه سبک می‌باشد به طوری که تعداد عملیات مَک (MAC[3]) آن کمتر از ۱۴۰,۰۰۰ و حافظه مصرفی آن کمتر از دو کیلوبایت است.

جدول ۲: ساختار شبکه عصبی اتصال کامل.

لایه	نوع لایه	تعداد نورون	تابع فعال‌سازی
۱	Dense	۵۳	ReLU
۲	Dense	۱۲۸	ReLU
۳	Dense	۲۵۶	ReLU
۴	Dense	۲۵۶	ReLU
۵	Dense	۱۲۸	ReLU
۶	Dense	۸	Softmax

جدول ۳ ساختار CNN پیشنهادی را توصیف می‌کند. این مدل دارای دو لایه پیچشی است که هرکدام ۳۲ فیلتر ۳ × ۳ دارند و به دنبالشان یک لایه نمونه‌برداری با اندازه پنجره ۲ × ۲ آمده است. پس از آنها دو لایه اتصال کامل با ۲۵۶ نورون با تابع فعال‌ساز ReLU قرار دارد که نرخ حذف تصادفی در زمان آموزش برای هرکدام ٪۵۰ است. در انتها نیز یک لایه اتصال کامل برای خروجی قرار دارد که دارای ۸ نورون با تابع فعال‌ساز Softmax است.

جدول ۳: شبکه عصبی پیچشی مورد استفاده در لایه محاسبه.

شماره	نوع لایه	فیلتر	پنجره	تعداد نورون	نرخ حذف تصادفی	تابع فعال‌سازی
۱	Conv2D	۳۲ × (۳, ۳)	–	–	–	ReLU
۲	MaxPooling2D	–	(۲, ۲)	–	–	–
۳	Conv2D	۳۲ × (۳, ۳)	–	–	–	ReLU
۴	MaxPooling2D	–	(۲, ۲)	–	–	–
۵	Dense	–	–	۲۵۶	۰/۵	ReLU
۶	Dense	–	–	۲۵۶	۰/۵	ReLU
۷	Dense	–	–	۸	–	Softmax

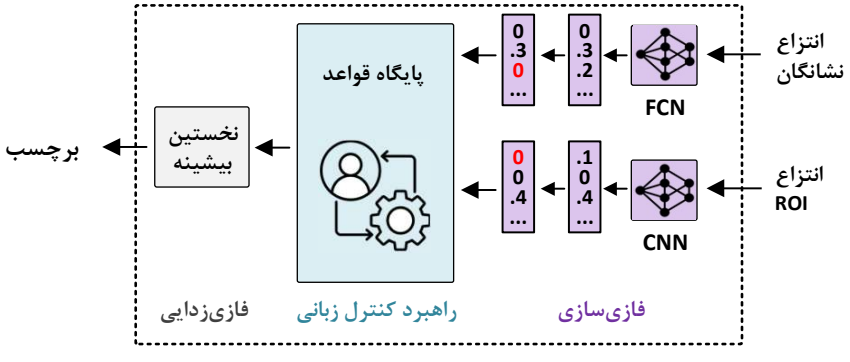
[1] Fully Connected Neural Network
[2] Shallow CNN
[3] Multiply–Accumulate

۳.۳. لایه همجوشی. در لایه سوم مدل پیشنهادی مبتنی‌بر ADFA، از یک مدل استنتاج فازی برای همجوشی تصمیم مدل‌ها استفاده می‌کنیم. این مدل، براساس سیستم استنتاج ممدانی[1] طراحی شده است که قابلیت مدل‌سازی دانش و تجربه افراد خبره در قالب پایگاه دانش و استفاده از آن در فرآیند استنتاج را دارد.

شکل ۶ نمای کلی رویکرد همجوشی خروجی مدل‌های محاسباتی را نمایش می‌دهد. مدل‌های پایه مورد استفاده، یعنی FCN و CNN، در لایه آخر خود از تابع فعال‌ساز Softmax استفاده می‌کنند و خروجی آنها بردار احتمالاتی به طول تعداد دسته‌ها است. بنابراین، یکی از ویژگی‌های بردار خروجی این است که حاصل جمع درایه‌های آن، عدد ثابت ۱ است. به عبارت دیگر، اگر تعداد کلاس‌ها را با N و احتمال تعلق ورودی به دسته i را با p_i نمایش دهیم، داریم:

$$\sum_{i=1}^{N} p_i = 1. \qquad (1.3)$$

قابل ذکر است که در لایه همجوشی، فرض بر این است که از هر مدل پایه موجود در لایه محاسبه، فقط یک بردار احتمال بعنوان تصمیم دریافت شود. از سوی دیگر، در اینجا داده ورودی یک ویدئو است که در روال زیرنمونه‌برداری، تعدادی از فریم‌ها برداشت شده و سپس برای هر فریم یک بردار احتمال خروجی توسط هر مدل پایه تولید می‌شود. بنابراین، پیش از همجوشی تصمیم مدل‌ها، بردارهای احتمال هر مدل، به روش میانگین‌گیری تجمیع می‌شود تا تصمیم مدل پایه، براساس یک فریم نباشد. در نتیجه، اطلاعات چندین فریم از ویدئو ورودی در بردار احتمال تولید شده برای هر مدل پایه تعبیه شده است.



شکل ۶: سیستم استنتاج فازی حاصل از ارتباط لایه‌های محاسبه و همجوشی.

[1] Mamdani Inference System

در سیستم استنتاج فازی شکل ۶، هر مدل پایه به‌عنوان یک متغیر زبانی در نظر گرفته شده که مقادیر زبانی آن، انواع احساسات (عصبانیت، تحقیر، تنفر، لذت، ترس، خنثی، غم و شگفتی) هستند. به عبارت دیگر، بردار احتمالاتی خروجی هر مدل پایه را می‌توان اینگونه تفسیر کرد که میزان شباهت ورودی به داده‌های آموزش هر احساس را نشان می‌دهد. این دیدگاه، یعنی تفسیر بردار خروجی به عنوان بردار شباهت ورودی به داده‌های دسته‌های مختلف، یک تفسیر فازی است. به بیان دیگر، مدل محاسباتی مانند یک تابع عضویت چندگانه عمل می‌کند که درجه شباهت ورودی به داده‌های آموزشی هر دسته را به صورت جداگانه محاسبه می‌کند. این تفسیر، امکان بهره‌گیری از منطق فازی برای همجوشی تصمیم مدل‌ها را فراهم می‌کند.

در مرحله فازی‌سازی، درجه عضویت ورودی در مجموعه فازی هرکدام از این مقادیر زبانی، مستقیماً توسط خروجی تجمیع شده مدل مربوطه (بردار Softmax) تأمین می‌شود. به بیان دیگر، توابع عضویت صریح برای متغیرهای ورودی تعریف نشده‌اند، بلکه مقادیر عضویت به‌صورت عددی از مدل‌های یادگیری ماشین به‌دست می‌آیند.

پس از دریافت درجه‌های عضویت فازی ورودی‌ها در دسته‌ها، ابتدا روی آنها αـ برش[1] اعمال می‌شود تا درجه عضویت‌های کمتر از α، صفر شوند. این عمل مقادیر نویز و غیرمعتبر را حذف می‌کند و تنها مقادیر با درجه عضویت بالا را در فرآیند استنتاج نگه می‌دارد؛ که باعث می‌شود نتایج نهایی دقیق‌تر و پایدارتر شوند. به عبارت دیگر، هدف از αـ برش، تمرکز بیشتر بر اطلاعات مهم و کاهش تأثیر مقادیر غیرمهم است. در این پژوهش، مقدار α برای هر بردار را برابر با دومین بیشینه آن قرار داده‌ایم.

در ادامه، با اعمال قواعد استنتاج، درجه‌های فازی خروجی محاسبه می‌شوند و سپس با فازی‌زدایی، خروجی نهایی به دست می‌آید. جدول ۴ پایگاه قواعد مربوط به راهبرد کنترل زبانی برای همجوشی تصمیم مدل‌ها را نمایش می‌دهد. در این جدول، مجموعه U شامل احساسات عصبانیت، غم، خنثی (بدون احساس)، ترس، شگفتی، لذت، تنفر و تحقیر است. تعداد قوانین مورد استفاده ۶۴ مورد است که به صورت فشرده در قالب ۱۵ قانون بیان شده‌اند. در اینجا، منظور از بیان فشرده این است که در هر ردیف، چند قاعده با هم بیان شده‌اند. همچنین، ممکن است چندین قاعده هم‌زمان فعال شوند و خروجی‌های متفاوت بدهند. این خروجی‌ها تجمیع شده و فازی‌زدایی اعمال می‌شود تا خروجی نهایی، یعنی برچسب احساسات، به دست آید. همچنین، در این سیستم استنتاج از روش نخستین بیشینه[2] برای فازی‌زدایی استفاده شده است.

---

[1] α-cut

[2] First Maxima

جدول ۴: پایگاه قواعد سیستم استنتاج فازی مورد استفاده در همجوشی تصمیم.

ردیف	مدل FCN	عملگر	مدل CNN	آنگاه
۱	عصبانیت	یا	عصبانیت	عصبانیت
۲	غم	و	$x \in U/\{$عصبانیت$\}$	غم
۳	$x \in U/\{$عصبانیت$\}$	و	غم	غم
۴	خنثی	و	$x \in U/\{$عصبانیت ، غم$\}$	خنثی
۵	$x \in U/\{$عصبانیت ، غم$\}$	و	خنثی	خنثی
۶	ترس	و	شگفتی	ترس
۷	شگفتی	و	ترس	شگفتی
۸	ترس	و	$x \in \{$ترس، لذت ، تنفر ، تحقیر$\}$	ترس
۹	شگفتی	و	$x \in \{$شگفتی، لذت ، تنفر ، تحقیر$\}$	شگفتی
۱۰	$x \in \{$لذت ، تنفر ، تحقیر$\}$	و	ترس	ترس
۱۱	$x \in \{$لذت ، تنفر ، تحقیر$\}$	و	شگفتی	شگفتی
۱۲	لذت	و	$x \in \{$لذت، تنفر ، تحقیر$\}$	لذت
۱۳	$x \in \{$تنفر ، تحقیر$\}$	و	لذت	لذت
۱۴	تنفر	و	$x \in \{$تنفر ، تحقیر$\}$	تنفر
۱۵	تحقیر	و	$x \in \{$تنفر ، تحقیر$\}$	تحقیر

مثال ۱.۳. فرض کنیم پس از لایه‌های انتزاع و محاسبه، خروجی مدل‌ها چنین باشد:

FCN : $\frac{۰/۴۰}{\text{عصبانیت}}$ ، $\frac{۰/۰۵}{\text{تحقیر}}$ ، $\frac{۰/۱۰}{\text{تنفر}}$ ، $\frac{۰/۳۵}{\text{ترس}}$ ، $\frac{۰/۱۰}{\text{شگفتی}}$ و **CNN :** $\frac{۰/۰۷}{\text{عصبانیت}}$ ، $\frac{۰/۰۳}{\text{تحقیر}}$ ، $\frac{۰/۲۰}{\text{تنفر}}$ ، $\frac{۰/۲۰}{\text{ترس}}$ ، $\frac{۰/۵۰}{\text{غم}}$.

پس از برش آلفا با $\alpha = ۰/۲$، درجه‌های کوچکتر از این مقدار حذف می‌شوند و داریم:

FCN : $\frac{۰/۴۰}{\text{عصبانیت}}$ ، $\frac{۰/۳۵}{\text{ترس}}$ و **CNN :** $\frac{۰/۲۰}{\text{تنفر}}$ ، $\frac{۰/۲۰}{\text{ترس}}$ ، $\frac{۰/۵۰}{\text{غم}}$.

آنگاه، با در نظر گرفتن عملگرهای $\vee$ و $\wedge$ که به ترتیب معادل "یا" و "و" در جدول ۴ هستند، با اعمال قواعد این جدول، خروجی‌های زیر به دست می‌آیند:

FCN: عصبانیت (۰/۴۰) $\vee$ **CNN: غم (۰/۵۰)** $\xLeftarrow{\text{قاعده ۱}}$ **خروجی: عصبانیت (۰/۵۰)**

FCN: عصبانیت (۰/۴۰) $\vee$ **CNN: ترس (۰/۲۰)** $\xLeftarrow{\text{قاعده ۱}}$ **خروجی: عصبانیت (۰/۴۰)**

FCN: عصبانیت (۰/۴۰) $\vee$ **CNN: تنفر (۰/۲۰)** $\xLeftarrow{\text{قاعده ۱}}$ **خروجی: عصبانیت (۰/۴۰)**

FCN: ترس (۰/۳۵) $\wedge$ **CNN: غم (۰/۵۰)** $\xLeftarrow{\text{قاعده ۳}}$ **خروجی: غم (۰/۳۵)**

FCN: ترس (۰/۳۵) $\wedge$ **CNN: ترس (۰/۲۰)** $\xLeftarrow{\text{قاعده ۸}}$ **خروجی: ترس (۰/۲۰)**

FCN: ترس (۰/۳۵) $\wedge$ **CNN: تنفر (۰/۲۰)** $\xLeftarrow{\text{قاعده ۸}}$ **خروجی: ترس (۰/۲۰)**

در گام بعد، روش فازی‌زدایی اولین‌بیشینه، روی خروجی‌های فازی اعمال می‌شود و با توجه به اینکه بیشترین درجه خروجی قواعد متعلق به **عصبانیت (۰/۵۰)** است، بنابراین تشخیص نهایی نیز **عصبانیت** خواهد بود.

۴. نتایج تجربی و ارزیابی

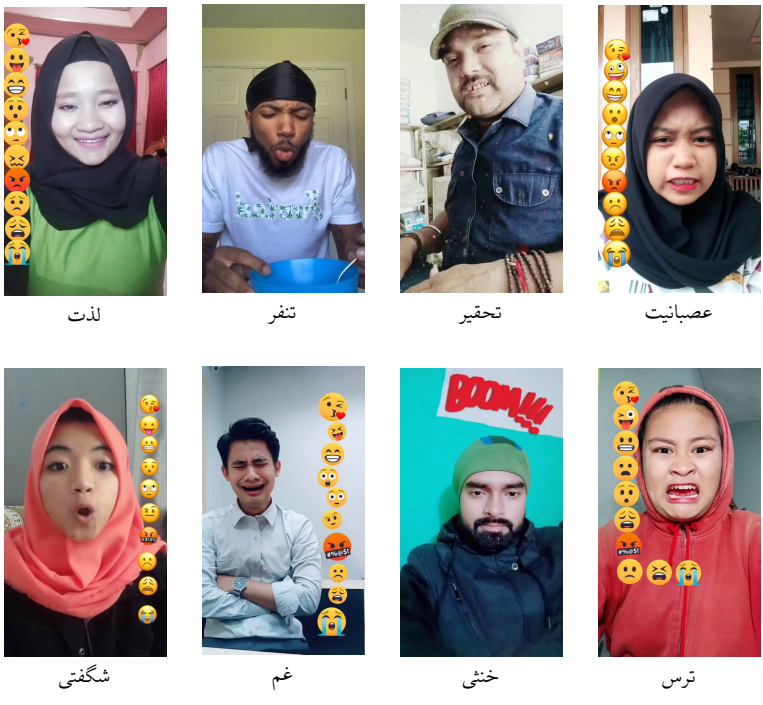
۱.۴. داده‌های آزمایش. برای نشان دادن کارآیی رویکرد پیشنهادی در کلاس‌بندی ویدئوهای کوتاه براساس احساسات افراد، از مجموعه داده‌های L-SVD [۲۱] استفاده شده است. این مجموعه داده در سال ۲۰۲۳ ارائه شده و پس از حذف داده‌های خراب، شامل بیش از ۱۴,۰۰۰ ویدئو کوتاه از افراد با جنسیت و نژادهای مختلف است. داده‌های مجموعه L-SVD در هشت کلاس عاطفی عصبانیت، تحقیر، تنفر، لذت، ترس، خنثی، غم و شگفتی طبقه‌بندی شده‌اند.

جدول ۵ مشخصات داده‌های هر کلاس مانند تعداد ویدئو، حجم میانگین، بیشترین تعداد فریم، کمترین تعداد فریم و میانگین تعداد فریم را نشان می‌دهد. براساس این اطلاعات، کلاس‌ها متوازن نیستند به طوری که کلاس خنثی با ۴۲۱۷ ویدئو دارای بیشترین قلم داده و کلاس ترس با ۳۵۰ ویدئو کوچک‌ترین کلاس است. این مسئله می‌تواند به نقص در تعمیم مدل در مواجهه با داده‌های جدید و درنتیجه دقت پایین‌تر در پیش‌بینی کلاس‌های کمتر شود. بنابراین، در کلاس‌های کوچکتر، نرخ زیرنمونه‌برداری را افزایش دادیم و کاستی تا حدودی جبران شد. در هر کلاس، ویدئوهای کوتاه از ۱۰ تا ۱۷ فریم و ویدئوهای بلند از ۱۰۲ تا ۱۴۴ فریم دارند.

جدول ۵: ویژگی‌های مجموعه داده L-SVD پس از حذف داده‌های خراب.

شماره	کلاس	تعداد ویدئو	حجم میانگین (kB)	بیشترین فریم	کمترین فریم	میانگین فریم
۰	عصبانیت	۱۴۹۷	۲۲۹	۱۴۴	۱۵	۴۲
۱	تحقیر	۱۰۲۸	۲۲۰	۱۳۰	۱۰	۳۹
۲	تنفر	۴۹۴	۲۰۰	۱۰۲	۱۱	۳۸
۳	لذت	۳۷۵۰	۲۴۴	۱۳۵	۱۰	۴۲
۴	ترس	۳۵۰	۱۸۶	۱۱۱	۱۴	۳۶
۵	خنثی	۴۲۱۷	۲۴۱	۱۲۵	۱۶	۴۴
۶	غم	۲۱۷۷	۲۲۵	۱۲۵	۱۷	۴۳
۷	شگفتی	۵۱۷	۱۸۹	۱۱۸	۱۲	۳۵

شکل ۷ نمونه داده‌های هر هشت کلاس را در قالب یک فریم منتخب از یک ویدئو نشان می‌دهد. اندازه فریم تمام ویدئوها برابر با ۱۰۲۴ × ۷۵۶ است. با توجه به نمونه‌های ارائه شده در این شکل، تنوع چهره، نژاد، جنسیت و پوشش افراد نمایان است. بعلاوه، پس‌زمینه تصاویر نشان می‌دهد که ویدئوها در محیط‌های واقعی و متنوع تهیه شده‌اند و برای آموزش مدل‌های یادگیری ماشین جهت استفاده در کاربردهای واقعی مناسب هستند. قابل ذکر است که تشخیص احساسات در این داده‌ها را فقط با پردازش حالت چهره انجام می‌دهیم.



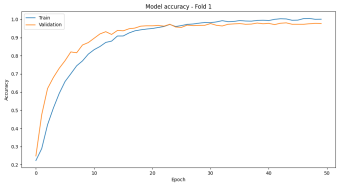
شکل ۷: یک فریم از نمونه داده‌های ویدئویی کلاس‌های مختلف.

۲.۴. ارزیابی مدل پیشنهادی. جدول ۶ مشخصات مدل‌های پایه FCN و CNN و مدل مبتنی‌بر ADFA حاصل از همجوشی آنها با سیستم استنتاج فازی را نمایش می‌دهد. مدل FCN در مقایسه با مدل CNN بسیار سبک است اما دقت آن حدود ٪۴ کمتر است. البته مدل CNN پیشنهادی نیز کم‌عمق و سبک است و میزان محاسبات و مصرف حافظه آن در مقایسه با CNN های عمیق ناچیز است. همجوشی این دو مدل پایه در معماری ADFA منجر به افزایش دقت به ٪۹۶/۷۲ شده در حالی که تعداد عملیات مَک مدل نهایی برای هر تشخیص، کمتر از ۱۴ میلیون و حجم آن حدود ۲۰۵ کیلو بایت است.

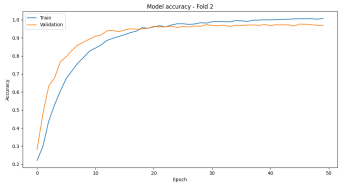
جدول ۶: مشخصات مدل‌های پایه و مدل مبتنی‌بر ADFA حاصل شده.

مدل	مَک (میلیون)	اندازه (کیلوبایت)	دقت %
FCN	۰/۱۳۹	۲	۸۹/۵۰
CNN	۱۳/۵۹۱	۲۰۳	۹۳/۴۲
ADFA	۱۳/۷۳۰	۲۰۵	۹۶/۷۲

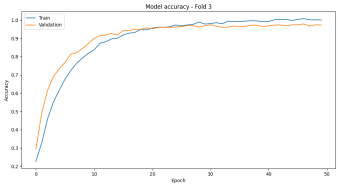
برای ارزیابی مدل مبتنی‌بر ADFA پیشنهادی، ابتدا اعتبارسنجی‌متقابل 5-Fold انجام شد که نتایج در شکل ۸ نمایش داده شده است. بر اساس این نمودارها، روند تغییرات دقت آموزش و اعتبارسنجی در تمام تکرارها الگوی مشابهی دارد؛ در ابتدا دقت اعتبارسنجی کمی بیشتر از دقت آموزش است که ناشی از اثرات منظم‌کننده یا وزن‌دهی اولیه مناسب مدل است. اما پس از چند دوره آموزش، دقت اعتبارسنجی به‌تدریج از دقت آموزش کمتر می‌شود و در نهایت هر دو به مقادیر مشابه و نسبتاً پایدار همگرا می‌شوند مقادیر نهایی برای میانگین دقت در بین پنج تکرار در بازه نسبتاً پایدار ٪۹۶/۸۷ تا ٪۹۷/۷۲ قرار دارند که تأییدی بر سازگاری عملکرد مدل در داده‌های دیده‌نشده است. این رفتار همگرا و پایدار در تمام تکرارها نشان می‌دهد که مدل پیشنهادی از نظر تعمیم‌پذیری عملکرد خوبی دارد.



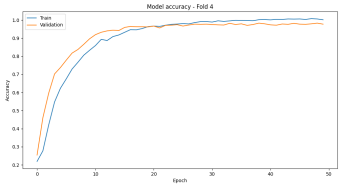
تکرار ۱: صحت: ۹۸/۸۱ بازیابی: ۹۶/۸۷
F1-score: ۹۷/۸۳ میانگین‌دقت: ۹۷/۷۲.



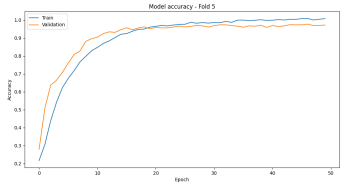
تکرار ۲: صحت: ۹۷/۶۲ بازیابی: ۹۶/۴۹
F1-score: ۹۷/۰۵ میانگین‌دقت: ۹۶/۸۷.



تکرار ۳: صحت :۹۷/۷۵ بازیابی: ۹۶/۹۶
F1-score: ۹۷/۳۵ میانگین‌دقت: ۹۷/۳۴.



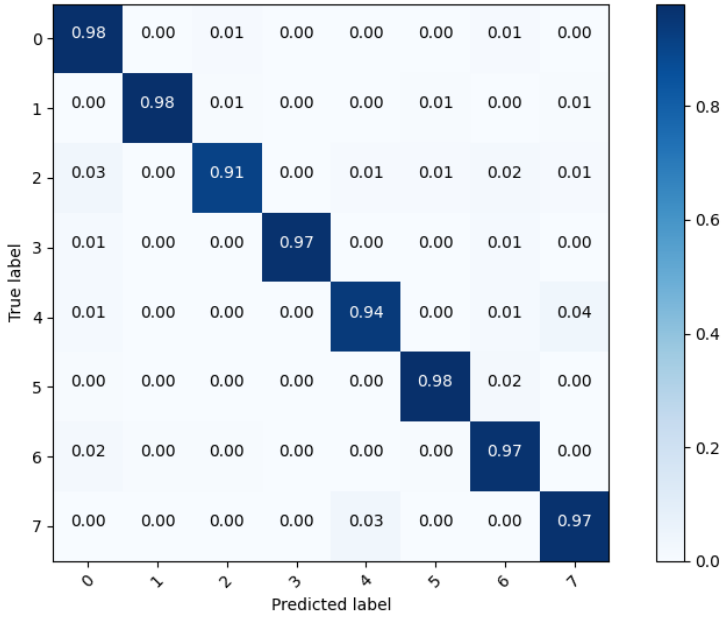
تکرار ۴: صحت: ۹۸/۸۴ بازیابی: ۹۷/۳۴
F1-score: ۹۸/۰۹ میانگین‌دقت: ۹۷/۷۲.



تکرار ۵: صحت: ۹۸/۰۸ بازیابی: ۹۶/۸۶
F1-score: ۹۷/۴۷ میانگین‌دقت: ۹۷/۲۴.

شکل ۸: نتایج اعتبارسنجی متقابل مدل پیشنهادی مبتنی‌بر ADFA.

شکل ۹ ماتریس درهم‌ریختگی ارزیابی مدل پیشنهادی را نشان می‌دهد و برچسب کلاس‌ها مطابق جدول ۵ است. براساس اطلاعات این ماتریس، دقت تشخیص در کلاس‌های مختلف از ٪۹۱ تا ٪۹۸ متغیر است. البته، فقط دو کلاس تنفر و ترس دقت پایین‌تر از ٪۹۷ نشان داده‌اند به طوری که احساس تنفر، در موارد اندک تشخیص اشتباه، بیشتر با عصبانیت یا غم اشتباه گرفته شده و احساس ترس نیز در مواردی به عنوان احساس شگفتی تشخیص داده شده است. با توجه به اطلاعات جدول ۵، به نظر می‌رسد که کم بودن تعداد داده‌های این دو کلاس تأثیر منفی اندکی روی آموزش مدل داشته و افزایش نرخ نمونه‌برداری تا حدود زیادی عدم توازن داده‌ها را جبران کرده است. همچنین، کلاس شگفتی هم بیشترین سردرگمی را با کلاس ترس داشته که به نظر می‌رسد مشترک بودن ویژگی‌هایی مانند باز بودن هم‌زمان دهان و چشم‌ها تأثیر اصلی را داشته‌اند. نکته دیگر در مورد کلاس‌های تنفر، ترس و شگفتی این است که تعداد نمونه‌های آموزشی آنها در مقایسه با سایر کلاس‌ها چند برابر کمتر است که باعث می‌شود فیلترهای مدل CNN در درک ویژگی‌های متمایز کننده این کلاس‌ها کمی ضعیف‌تر باشند. البته از نظر انسانی هم تشخیص این دو احساس از روی تصاویر یا ویدئو در مواردی بسیار دشوار و گاهی غیرممکن می‌شود.



شکل ۹: ماتریس درهم‌ریختگی مدل پیشنهادی مبتنی‌بر ADFA.

۳.۴. مقایسه مدل پیشنهادی با مدل‌های یادگیری ماشین مدرن. جدول ۷ مشخصات مدل پیشنهادی مبتنی بر معماری ADFA و چهار مدل مدرن برای تشخیص احساسات را نمایش می‌دهد. در این جدول، مدل‌ها بر اساس تعداد عملیات مَک مرتب شده‌اند. نتایج نشان می‌دهد علاوه بر اینکه مدل پیشنهادی دارای بالاترین دقت در میان این مدل‌ها است، تعداد عملیات مَک و اندازه آن نیز به شکل قابل توجهی کمتر است. این اطلاعات نشان دهنده کارآیی مدل پیشنهادی در تشخیص احساسات از داده‌های ویدئویی است. بعلاوه، با توجه به سبک بودن این مدل، برای کاربردهای منابع‌محور مانند مانیتورینگ سلامت، تشخیص فعالیت‌های فیزیکی و سایر کاربردهای مرتبط با اینترنت اشیاء که نیازمند ارائه نتایج دقیق و سرعت پردازش بالا است، پتانسیل استفاده بالایی دارد.

جدول ۷: مقایسه مدل پیشنهادی با مدل‌های مدرن تشخیص احساسات.

مدل	مَک (میلیون)	اندازه (مگابایت)	دقت %
مدل پیشنهادی مبتنی بر معماری ADFA	۱۳/۷۳۰	۰/۲۰۰	۹۶/۷۲
ResNet-18 [۹]	۲۶۱/۲۶	۱۵/۹۴	۸۸/۷۶
ترکیب TLF-ResNet-18 و SVM [۸]	۳۰۰	۵۰	۹۳/۴۴
ترکیب ResNet-50 و DenseNet-121 [۱۰]	۶۷۷/۴۶	۱۸۷/۵	۹۱/۶۹
CNN [۱۹]	۱۷۵۰	۴۱/۱۴	۹۳/۷۰

۵. نتیجه‌گیری

در این پژوهش، یک مدل یادگیری ماشین مبتنی بر معماری ADFA برای تشخیص احساسات در داده‌های ویدئویی توسعه داده شد. در این راستا، ابتدا دو روال انتزاع داده در لایه نخست تعریف شد که هر دو انتزاع با دیدگاه‌های مختلف باعث کاهش حجم و پیچیدگی داده‌های ورودی شده‌اند. در ادامه، دو مدل سبک وزن شبکه اتصال کامل و شبکه پیچشی کم‌عمق در لایه دوم تعریف شد که هر کدام از داده‌های یک انتزاع برای تصمیم‌گیری استفاده می‌کنند. سپس، خروجی این مدل‌های پایه با استفاده از یک سیستم استنتاج فازی ممدانی همجوشی داده شد تا مدل پیشنهادی حاصل شود. ارزیابی مدل پیشنهادی با آموزش و تست روی مجموعه داده‌های ویدئویی L-SVD انجام شد که شامل هشت کلاس مختلف احساسات است. نتایج اعتبارسنجی و همچنین تست مدل روی این مجموعه نشان‌دهنده دقت بیش از ٪۹۶ است در حالی که این مدل برای تشخیص احساسات از داده‌های ویدئویی با تکنیک زیرنمونه‌برداری، میزان پردازش اندکی

دارد و برای درک هر فریم، کمتر از ۱۴ میلیون عملیات مَک را انجام می‌دهد. این ویژگی برای کاربردهای تشخیص احساسات در زمان واقعی از اهمیت ویژه برخوردار است. نتایج حاصل از این پژوهش نشان‌دهنده پتانسیل بالای معماری ADFA و همچنین انتزاع‌ها و روش همجوشی پیشنهادی برای توسعه مدل‌های یادگیری ماشین درک ویدئو در محیط‌های دارای منابع محدود است. نوآوری‌های این پژوهش را می‌توان در چند محور کلیدی خلاصه کرد: برای نخستین‌بار معماری ADFA در تحلیل داده‌های ویدئویی به‌کار گرفته شده است که امکان پردازش بلادرنگ و سازگار با محیط‌های منابع‌محور را فراهم می‌سازد. در این چارچوب، مدلی سبک‌وزن و در عین حال دقیق برای تشخیص احساسات توسعه یافته و برای بهبود کارایی، دو نوع انتزاع به‌منظور خلاصه‌سازی داده‌های ویدئویی تعریف شده است. در نهایت، یک سیستم استنتاج فازی ممدانی طراحی شده که نقش مهمی در همجوشی خروجی مدل‌های یادگیری ماشین و ارتقای دقت کلی ایفا می‌کند.

با توجه به عملکرد موفق روش پیشنهادی، قصد داریم در مطالعات آتی قابلیت تعمیم‌پذیری معماری ADFA به سایر حوزه‌های تحلیل ویدئو، مانند تشخیص حرکات بدن و پایش وضعیت جسمی کاربران را مورد بررسی قرار دهیم. همچنین، به دنبال توسعه انتزاع‌های مبتنی‌بر توجه (Attention) و ادغام آن‌ها با مدل‌های حافظه‌دار LSTM و Transformer برای افزایش توانایی تشخیص الگوهای زمانی پیچیده‌تر خواهیم بود.

مراجع

[۱] صفائی، محمد ، مغاری، سمیه ، فلاح، محمد کاظم و غزنوی، مهرداد . (۱۴۰۳). به کارگیری فرآیند تحلیل سلسه‌مراتبی فازی برای گزینش مدل‌های توسعه داده شده با معماری انتزاع و همجوشی تصمیم (مورد مطالعه: دسته‌بندی حروف دستنویس فارسی). تصمیم گیری و تحقیق در عملیات، ۱۴۰۳، ـ. doi: ۱۸۶۳.۴۷۳۲۱۷.۲۰۲۴.dmor/۲۲۱۰۵.۱۰

[2] Ahmed, Naveed, Zaher Al Aghbari, and Shini Girija. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications, 17, 200171.

[3] Bai, Zhongli, Junjie Liu, Fazheng Hou, Yirui Chen, Meiyi Cheng, Zemin Mao, Yu Song, and Qiang Gao. (2023). Emotion recognition with residual network driven by spatial-frequency characteristics of EEG recorded from hearing-impaired adults in response to video clips. Computers in Biology and Medicine, 152, 106344.

[4] Chen, Jin, Tony Ro, and Zhigang Zhu. (2022). Emotion recognition with audio, video, EEG, and EMG: a dataset and baseline approaches. IEEE Access, 10, 13229-13242.

[5] Chen, Yen-Liang, Chia-Ling Chang, and Chin-Sheng Yeh. (2017). Emotion classification of YouTube videos. Decision Support Systems, 101, 40-50.

[6] Fallah, Mohammad K., Najafi, Mohammadreza, Gorgin, Saeid, and Lee, Jeong-A. (2024). Abstraction and decision fusion architecture for resource-aware image understanding with application on handwriting character classification. Applied Soft Computing, 2024, 111813.

[7] Fallah, Mohammad K., Najafi, Mohammadreza, Gorgin, Saeid, and Lee, Jeong-A. (2024). An ultra-low-computation model for understanding sign languages. Expert Systems with Applications, 249, 123782.

[8] Haider, Irfan, Hyung-Jeong Yang, Guee-Sang Lee, and Soo-Hyung Kim. (2023). Robust human face emotion classification using triplet-loss-based deep CNN features and SVM. Sensors, 23(10), 4770.

[9] Huang, Yahe, and Dongying Bo. (2023). Emotion classification and achievement of students in distance learning based on the knowledge state model. Sustainability, 15(3), 2367.

[10] Khanna, Deepanshu, Neeru Jindal, Prashant Singh Rana, and Harpreet Singh. (2024). Enhanced spatio-temporal 3D CNN for facial expression classification in videos. Multimedia Tools and Applications, 83(4), 9911-9928.

[11] Karatay, Buşra, Deniz Beştepe, Kashfia Sailunaz, Tansel Özyer, and Reda Alhajj. (2024). CNN-Transformer based emotion classification from facial expressions and body gestures. Multimedia Tools and Applications, 83(8), 23129-23171.

[12] Khare, Smith K., Victoria Blanes-Vidal, Esmaeil S. Nadimi, and U. Rajendra Acharya. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Information fusion, 102, 102019.

[13] Lek, Jeniffer Xin-Ying, and Jason Teo. (2023). Academic emotion classification using fer: A systematic review. Human Behavior and Emerging Technologies, 2023(1), 9790005.

[14] Loewenstern, Yocheved, Noa Benaroya-Milshtein, Katya Belelovsky, and Izhar Bar-Gad. (2024). Automatic Identification of Facial Tics Using Selfie-Video. IEEE Journal of Biomedical and Health Informatics.

[15] Ma, Zhuang, Ao Li, Jiehao Tang, Jianhua Zhang, and Zhong Yin. (2025). Multimodal emotion recognition by fusing complementary patterns from central to peripheral neurophysiological signals across feature domains. Engineering Applications of Artificial Intelligence, 143, 110004.

[16] Miranda Calero, Jose A., Laura Gutiérrez-Martín, Esther Rituerto-González, Elena Romero-Perales, Jose M. Lanza-Gutiérrez, Carmen Peláez-Moreno, and Celia López-Ongil. (2024). WEMAC: Women and emotion multi-modal affective computing dataset. Scientific Data, 11(1), 1182.

[17] Moontaha, Sidratul, Franziska Elisabeth Friederike Schumann, and Bert Arnrich. (2023). Online learning for wearable EEG-based emotion classification. Sensors, 23(5), 2387.

[18] Nguyen, Van-Ho, Nghia Nguyen, Thuy-Hien Nguyen, Yen-Nhi Nguyen, Mai-Thu Dinh, and Dung Doan. (2025). Customer emotion detection and analytics in hotel and tourism services using multi-label classificational models based on ensemble learning. Annals of Operations Research, 1-31.

[19] Pandeya, Yagya Raj, and Joonwhoan Lee. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools and Applications, 80(2), 2887-2905.

[20] Pei, Guanxiong, Qian Shang, Shizhen Hua, Taihao Li, and Jia Jin (2024). EEG-based affective computing in virtual reality with a balancing of the computational efficiency and recognition accuracy. Computers in Human Behavior, 152, 108085.

[21] Peiran L, Linbo T, Xizheng Y. University of Wisconsin Madison. (2023) L-SVD: A Comprehensive Video Dataset for Emotion Recognition, Available: https://github.com/PeiranLi0930/emotionnet.

[22] Radzi, Nor Haizan Mohamed, and Haslina Hashim. (2024). Research on Emotion Classification Based on Multi-modal Fusion. Baghdad Science Journal, 21(2 (SI)), 0548-0548.

[23] Song, Peipei, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang. (2024). Emotional video captioning with vision-based emotion interpretation network. IEEE Transactions on Image Processing.

[24] Wei, Jie, Guanyu Hu, Xinyu Yang, Anh Tuan Luu, and Yizhuo Dong. (2024). Learning facial expression and body gesture visual information for video emotion recognition. Expert Systems with Applications, 237, 121419.