Electronic & Cyber Defense
Vol. 12, No. 4, Winter 2025, Serial No. 48, PP.21-32
ISSN: 2322-4347, E-ISSN: 2980-8979

A hybrid model based on CNN-LSTM for speech emotion reognition

R. Ahmadian*, H. Rayatparvar A.Sarkardehee

*Researcher, Imam Hossein comprehensive University, Tehran, Iran

(Received: 2024/06/12, Revised: 2024/12/27, Accepted: 2025/01/12, Published: 2025/02/01)
DOR: https://dorl.net/dor/20.1001.1.23224347.1402.11.2.1.4

ABSTRACT

The data published in the virtual space, including text, image, video and speech, have become a reliable source for measuring the thoughts, opinions and emotions of the audience towards various objects such as governments, policies, personalities, products, etc. In order to conflict with the cognitive threats of the cyberspace, it is very important to recognition the cognitive structure of insider and enemy audiences. The current research was conducted in order to present a computational model for speech emotion recognition based on the combination of two Classifier of CNN-LSTM. In this article, at the beginning, the introductions about speech emotion recognition and its applications are mentioned, then the papers presented in the authoritative journals are reviewed and their accuracy is evaluated, in the following, a practical method for recognition the eight basic emotions of the audience including happiness, sadness, fear, calm, anger, Disgust, surprise, and neutral are presented. In this research, in order to have a high number of data, by combining the two data sets RAVDESS and TESS, a general data set was collected, in the feature extraction phase, three features MFCC, MEL and ZCR were extracted and combined, and then in the model designed by the combination of CNN and LSTM classifiers for Training and testing have been used. With the evaluations, the accuracy of the model on the test data is 92.57%, which is more accurate than the existing models.

Keywords: speech emotion recognition, audience evaluation, convolutional neural network, recurrent neural network, mel frequency capstral coefficients.

Cite this article: M. Yousefvand and M. Nasiri and R. Bibak “Improved implementation of image processing algorithm with HLS software for use in optical seeker,” Electronic and Cyber Defense, vol.12 , no.4 , pp.21-32 , .
DOR: https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.4.2.6
 Publisher: Imam Hossein University

*Corresponding Author Email: aboulfazlsarkardei@gmail.com

"پدافند الکترونیکی و سایبری"
سال دوازدهم، شماره ۴، زمستان ۱۴۰۳، ص ۳۲-۲۱
شاپا الکترونیکی: 2980-8979 شاپا چاپی: 2322-4347

علمی - پژوهشی

ارائه مدلی ترکیبی مبتنی بر CNN–LSTM جهت تشخیص هیجان از سیگنال گفتار

رضا احمدیان[۱]، حسین رعیت‌پرور[۲]، ابوالفضل سرکردەئی[۳*]

۱- کارشناسی ارشد، ۲- کارشناسی ارشد، ۳- پژوهشگر، دانشگاه امام حسین (ع)، تهران، ایران

(دریافت: ۱۴۰۳/۰۳/۲۳، بازنگری: ۱۴۰۳/۱۰/۰۷، پذیرش: ۱۴۰۳/۱۰/۲۳، انتشار: ۱۴۰۳/۱۱/۱۳)

DOR: https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.4.3.7

چکیده

داده‌های منتشرشده در فضای مجازی شامل متن، تصویر، ویدئو و صوت به منبعی معتبر برای سنجش افکار، عقاید و هیجانات مخاطب نسبت به اشیا مختلف مانند دولت‌ها، سیاست‌ها، شخصیت‌ها، محصولات و غیره تبدیل‌شده‌اند، به‌منظور مقابله با تهدیدات شناختی فضای سایبری، تشخیص شاکله شناختی مخاطبان خودی و غیرخودی بسیار حائز اهمیت است. پژوهش حاضر به‌منظور ارائه مدلی محاسباتی برای تشخیص هیجان گفتار مخاطب مبتنی بر ترکیب دوطبقه بند LSTM – CNN صورت گرفته است. در این مقاله در ابتدا مقدمه‌ای در مورد تشخیص هیجان گفتار و کاربردهای آن گفته‌شده، سپس طرح‌های ارائه‌شده در مجلات معتبر مرور و دقت آن‌ها ارزیابی‌شده است، در ادامه روشی کاربردی جهت تشخیص هشت هیجان پایه مخاطب شامل شادی، غم، ترس، آرام، خشم، نفرت، شگفت‌زده و خنثی ارائه‌شده است. در این پژوهش به‌منظور داشتن تعداد داده بالا، با ترکیب دو مجموعه داده RAVDESS و TESS یک مجموعه داده کلی جمع‌آوری‌شده، در مرحله استخراج ویژگی سه ویژگی MFCC، MEL و ZCR استخراج و ترکیب‌شده و سپس در مدل طراحی‌شده از ترکیب طبقه‌بندی کننده‌های CNN و LSTM جهت آموزش و آزمون استفاده‌شده است. با ارزیابی‌های انجام‌شده، دقت مدل بر روی‌داده‌های آزمون، ۹۲،۵۷ درصد است که نسبت به مدل‌های موجود دارای دقت بالاتری است.

کلیدواژه‌ها: تشخیص هیجان گفتار، مخاطب‌سنجی، شبکه عصبی کانولوشن، شبکه عصبی بازگشتی، ضرایب کپسترال فرکانس مل

۱. مقدمه

بیان هیجانات از طریق ابزارهای مختلفی مانند پاسخ‌ها، زبان، رفتار، حرکات بدن، وضعیت و غیره انجام می‌شود. بسیاری از فرایندهای فیزیولوژیکی مانند تنفس، ضربان قلب، دما، پوست و دما، پتانسیل‌های ماهیچه‌ای و غشای عصبی نیز می‌توانند برای بیان هیجانات مورداستفاده قرار بگیرند. [۱]. زبان انسانی تنها شامل اطلاعات لغوی نیست و درک کامل جملات بیان‌شده بین افراد مستلزم واقف بودن به‌تمامی ابعاد کلام است. هیجان یکی از اساسی‌ترین و مهم‌ترین این ابعاد است. در یک نگاه کلی اطلاعات گفتار شامل دو وجه زبانی یا معنای لغت‌ها و آوا یا نحوه بیان این لغات است. جان مارشال ریودر[۱] این دو وجه از زبان را به نام‌های دیدگاه شناختی و زیستی معرفی می‌کند و سپس با توجه به دیدگاه باک[۲] موضع دو سیستمی را در قبال عواطف ارجح می‌داند. سیستم اول، زیستی است و منشأ آن به تکامل گونه تاریخی برمی‌گردد. اطلاعات حسی به‌سرعت، خودکار و به‌صورت ناخودآگاه توسط سیستم لمبیک[۳] پردازش می‌شود. سیستم دوم شناختی است و به تاریخچه اجتماعی و فرهنگی منحصربه‌فرد انسان بستگی دارد. رویکرد بررسی عواطف با استفاده از اطلاعات آوایی گفتار نتیجه سیستم زیستی و تغییرات فیزیولوژیک حنجره است. عاطفه به‌عنوان یکی از پیچیده‌ترین حالات درونی و ذهنی انسان‌ها شناخته می‌شود. انسان‌ها برای توصیف عواطف خود بیشترین تلاش را انجام می‌دهند اما به نسبت سایر مفاهیم در ارتباطات، کمترین توفیق در انتقال منظور نتیجه می‌شود [۲]. مدل مختلفی از هیجانات به‌منظور مفهوم‌سازی و تمایز بین هیجانات تعریف می‌شوند که در زمان‌های

[1] Reeve.

[2] Buck.
[3] Limbic System.

استناد: احمدیان، رضا، رعیت‌پرور، حسین، سرکردەئی، ابوالفضل " ارائه مدلی ترکیبی مبتنی بر CNN – LSTM جهت تشخیص هیجان از سیگنال گفتار"، پدافند الکترونیکی و سایبری، ۱۲(۴)، ۳۲-۲۱، ۱۴۰۳. https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.4.3.7

ناشر: دانشگاه جام امام حسین(ع).

* رایانامه نویسنده مسئول: aboulfazlsarkardei@gmail.com

مختلف و توسط افراد مختلف، بیان‌شده است. نظریه هیجانات پایه، توسط اکمن [1] پیشنهادشده است که این هیجانات شامل خنثی، خشم، شادی، شگفت‌زده، نفرت، غم و ترس است. هیجان یکی از مؤلفه‌های مهم در سنجش شناختی و محاسباتی کاربر بر اساس داده‌هایی مانند متن، تصویر و فیلم و صوت و غیره است. با افزایش حجم داده‌های منتشرشده در فضای مجازی، این داده‌ها به منبعی معتبر برای سنجش افکار، عقاید، نگرش‌ها، احساسات و هیجانات کاربر نسبت به اشیا مختلف مانند دولت‌ها، سیاست‌ها، شخصیت‌ها، محصولات و غیره تبدیل‌شده‌اند. با توجه به سهولت و سرعت در ضبط و ارسال پیام‌های صوتی، گفتار به یکی از منابع اصلی انتقال پیام در شبکه‌های اجتماعی مختلف تبدیل‌شده و ارسال پیام‌های متنی جای خود را به گفتگوهای صوتی داده است. از یک سیگنال گفتاری، می‌توان مقادیر زیادی اطلاعات مانند جنسیت، کلمات، گویش، احساسات، سن و غیره را جمع‌آوری کرد که می‌توان از آن‌ها برای کاربردهای مختلف استفاده کرد. یکی از این کاربردها، تشخیص هیجانات گفتار است که در مطالعات به‌عنوان سامانه‌های SER [2] شناخته می‌شوند، است. سیستم تشخیص هیجان گفتار مبتنی بر یادگیری عمیق از بخش‌های زیر تشکیل می‌شود: بخش اول مجموعه داده که به‌منظور استفاده در قسمت آموزش و ارزیابی مدل استفاده می‌شود، مجموعه‌های داده بخشی ضروری از تشخیص هیجانات گفتار هستند زیرا فرآیند طبقه‌بندی بر داده‌های برچسب‌گذاری شده متکی است. کیفیت داده‌ها بر موفقیت فرآیند شناسایی تأثیر می‌گذارد. داده‌های ناقص، باکیفیت پایین یا معیوب ممکن است منجر به‌پیش بینی‌های نادرست شود. بنابراین، داده‌ها باید با دقت طراحی و جمع‌آوری شوند [۳].

بخش دوم شامل استخراج ویژگی است، ویژگی‌ها جنبه مهمی از تشخیص هیجانات گفتار هستند. مجموعه‌ای از ویژگی‌هایی که با دقت طراحی‌شده‌اند که با موفقیت هر هیجان را مشخص می‌کنند و نرخ تشخیص را افزایش می‌دهند. ویژگی‌های مختلفی برای سیستم‌های SER استفاده‌شده است. بااین‌حال، هیچ مجموعه‌ای از ویژگی‌های پذیرفته‌شده برای طبقه‌بندی دقیق و متمایز وجود ندارد. مطالعات موجود همگی تجربی بوده‌اند. گفتار یک سیگنال پیوسته با طول‌های مختلف است که حامل اطلاعات و هیجانات است. بنابراین، بسته به رویکرد موردنیاز، می‌توان ویژگی‌های سرا سری یا محلی را استخراج کرد و قسمت نهایی طراحی مدل تشخیص هیجان گفتار است که در ادامه موردبررسی قرار می‌گیرد [۴].

بخش سوم سیستم‌های تشخیص عواطف گفتار، هیجانات را برای یک گفتار معین طبقه‌بندی می‌کنند. علاوه بر طبقه‌بندی کننده‌های سنتی و الگوریتم‌های یادگیری عمیق، بسیاری از الگوریتم‌های یادگیری ماشین برای تشخیص هیجانات گفتار استفاده می‌شوند.

[1] Ekman.
[2] Speech emotion recognition.

بااین‌حال، درست مانند هر مشکل پیچیده‌ای، هیچ الگوریتم یادگیری ماشینی پذیرفته‌شده‌ای وجود ندارد که بتوان از آن استفاده کرد [۳]. یکی از طبقه بندها شبکه‌های عصبی کانولوشنال [3] (CNN) نوع خاصی از شبکه‌های عصبی هستند که برای پردازش داده‌هایی طراحی‌شده‌اند که دارای توپولوژی شبکه‌ای مانند تصاویر هستند. CNN با استفاده از چندین فیلتر مرتبط می‌تواند وابستگی‌های زمانی و مکانی را از یک منبع ورودی با موفقیت ثبت کند. ورودی‌ها به شکلی بدون از دست دادن ویژگی کاهش می‌یابند به‌طوری‌که پیچیدگی محاسباتی کاهش می‌یابد و میزان موفقیت الگوریتم افزایش می‌یابد. یک CNN از چندین لایه تشکیل‌شده است: لایه پیچشی، لایه پولینگ [4] و لایه کاملاً متصل [5] [۳].

۲. روش تحقیق

در این مقاله جهت تشخیص هیجان گفتار، ابتدا از دو مجموعه داده مختلف استفاده می‌شود، در ادامه از مجموعه داده نهایی، ویژگی‌ها استخراج می‌شود و سپس با استفاده از مدل طراحی‌شده، آموزش، آزمون و ارزیابی انجام می‌شود. در ادامه به کارهای مرتبط و روش کار انجام‌شده به‌صورت جزئی‌تر پرداخته می‌شود.

۲-۱. کارهای مرتبط

بین [6] و همکاران در پژوهشی به بررسی استفاده از طبقه‌بندی CNN در سامانه‌های تشخیص هیجان گفتار پرداختند، آن‌ها یک سیستم تشخیص هیجان گفتار مبتنی بر چهار طبقه بند CNN، RNN، SVM و MVR را به‌طور جداگانه ارائه و مقایسه کردند. مدلی که با مجموعه داده‌های SAVEE و با استفاده از ویژگی MFCC آموزش، طبقه‌بندی و آزمون شد. درنهایت، بهترین دقت نهایی برای CNN به دست آمد که ۸۳٫۶۱٪ بود [۵].

ماخمودوف [7] و همکاران در سال ۲۰۲۲ یک مدل جدید مبتنی بر CNN ارائه دادند که به‌طور موازی ویژگی‌های MFCC، فرازبانی و طیف نگار گفتاری را با طراحی معماری‌های مختلف CNN به‌صورت جداگانه برای ویژگی‌ها، استخراج و کدگذاری می‌کرد، علاوه بر این ویژگی‌های کدگذاری شده به سازوکارهای توجه برای نمایش بیشتر داده شدند و سپس طبقه‌بندی شدند. صحت تجربی روی مجموعه داده‌های باز DB-EMO و IEMOCAP اجرا شد و نتایج نشان داد که مدل پیشنهادی کارآمدتر از مدل‌های پایه است. WA و UA مدل پیشنهادی به ترتیب برابر با ۷۱٫۸ درصد و ۷۰٫۹ درصد در مجموعه داده DB-EMO بود. علاوه بر این، نرخ WA و

[3] Convolutional Neural Network.
[4] Polling.
[5] Fully-Connected.
[6] Bin.
[7] Makhmudov.

UA در مجموعه داده IEMOCAP ۷۲٫۴٪ و ۷۱٫۱٪ بود [۶].

پنگ[1] و همکاران در سال ۲۰۲۱ یک مدل مبتنی بر شبکه عصبی کانولوشنال مقیاسی بر روی مجموعه داده IEMOCAP با استفاده از ویژگی MFCC پیشنهاد دادند، آن‌ها از مکانیسم توجه جهت بهبود مدل خود استفاده کردند. نتایج ارزیابی دقت ۸۱٫۴ را نشان داد. [۷].

ژانگ[2] و همکاران یک روش جدید برای تشخیص هیجان گفتار مبتنی بر شبکه عصبی پیچش عمیق و مدل حافظه کوتاه‌مدت دو جهتِ پیشنهاد دادند. آن‌ها ابتدا نمونه‌های گفتار را با افزایش داده‌ها و متعادل کردن مجموعه داده‌ها پردازش کردند. سپس طیف‌نگاری‌های Mel سه کانالِ را به‌عنوان ورودی DCNN استخراج کردند. سپس مدل DCNN از پیش آموزش داده‌شده بر روی مجموعه داده ImageNet اعمال می‌شود. در مرحله بعد، BLSTM را به‌منظور یادگیری ویژگی‌های عاطفی سطح بالا و به دنبال آن‌یک لایه توجه که می‌تواند بر ویژگی‌های مرتبط بااحساس تمرکز کند، اتخاذ می‌شود. درنهایت، ویژگی‌های احساسی سطح بالا به شبکه عصبی عمیق برای پیش‌بینی هیجان نهایی استفاده می‌شوند. آزمایش‌ها بر روی پایگاه داده DB-EMO و IEMOCAP میانگین دقت وزن نشده را به ترتیب ۸۷٫۸۶ و ۶۸٫۵۰ درصد را به دست می‌دهد [۸].

تربج ور[3] و همکاران سیستمی را پیشنهاد کردند که CNN را با شبکه‌های LSTM ترکیب می‌کند که در آن از CNN به‌منظور یادگیری خودکار بهترین ویژگی‌های توصیفی سیگنال گفتار از طریق نمایش زمان خام استفاده می‌شود. سیگنال‌های گفتار ابتدا قطعه‌بندی می‌شوند، سپس نویز زدایی به‌عنوان مرحله پیش‌پردازش اعمال می‌شود. در ادامه، با استفاده از CNN، ویژگی‌های صوتی استخراج می‌شوند. درنهایت، RNN عمیق لایه‌ای LSTM با ویژگی‌های استخراج‌شده تغذیه می‌شود. عملکرد روش پیشنهادی به‌طور قابل‌توجهی بهتر از در مقایسه با ویژگی‌های طراحی‌شده سنتی است. رویکرد مشابهی توسط لیم[4] و همکاران اتخاذشده است. آن‌ها زمان پیشنهادی CNN توزیع‌شده را با CNN و RNN-LSTM مقایسه کردند. آن‌ها با استفاده از روش پیشنهادی دقت متوسط ۸۸٫۰۱٪ را به دست آوردند، درحالی‌که طبقه‌بندی CNN و LSTM به ترتیب میانگین دقت ۸۶٫۳۲٪ و ۷۸٫۳۱٪ را نتیجه دادند [۹, ۱۰].

وو[5] و همکاران از درخت متا تصمیم[6] برای طبقه‌بندی داده‌های گفتار و حداکثر سطح آنتروپی برای توصیف رابطه بین حالات عاطفی و قوانین انجمن هیجانی که از پایگاه دانش چینی استخراج‌شده‌اند، استفاده کردند، در ادامه از روش همجوشی محصول وزنی برای ترکیب هر دو نتیجه نهایی استفاده شد. روش‌های تشخیص مبتنی بر گفتار و روش‌های مبتنی بر برچسب معنایی به ترتیب دقت ۸۰٪ و ۸۰٫۹۲٪ را به دست آوردند. شناخت چندوجهی به نرخ شناسایی ۸۳٫۵۵ درصد دست‌یافت [۱۱].

در پژوهشی دیگر ایبن[7] و همکاران یک سیستم SER آنلاین با استفاده از RNN-LSTM پیشنهاد کردند. آن‌ها بعد زمان را به مدل هیجانی فعال‌سازی- ظرفیت دوبعدی اضافه کردند تا یک مدل سه‌بعدی جدید ایجاد کنند. انگیزه شامل افزودن بعد سوم زمانی به ارزیابی تدریجی هیجانات در زمان واقعی بود. ازآنجایی‌که LSTM-RNN به هیچ‌گونه تقسیم‌بندی علاوه بر فرم‌بندی سطح پایین نیاز ندارد، برای کاربردهای بلادرنگ مناسب است. عامل دیگری برای استفاده از RNN-LSTM این واقعیت است که مناسب بودن آن برای سری‌های زمانی متصل است. ویژگی‌های عروضی، طیفی و کیفیت صدا به‌عنوان ویژگی‌های صوتی استفاده می‌شود. علاوه بر این، ویژگی‌های زبانی استخراج‌شده و ویژگی‌های صوتی و زبانی با سطح ویژگی همجوشی ترکیب می‌شوند [۱۲].

تیان[8] و همکاران از همجوشی سلسله‌مراتبی برای ترکیب ویژگی‌های صوتی و واژگانی برای تشخیص هیجانات با مدل عاطفی استفاده کردند. آن‌ها از LSTM برای طبقه‌بندی با LLD، eGeMAPS و ویژگی‌های عروضی global به‌عنوان ویژگی‌های آکوستیک استفاده کردند و ویژگی‌های Disfluency و صداسازی غیرکلامی به‌عنوان ویژگی‌های واژگانی استفاده می‌شود. در همجوشی سلسله‌مراتبی، یک شبکه LSTM با سه لایه پنهان پیشنهادشده است. در لایه اول از ویژگی‌های LLD و eGeMAPS استفاده می‌شود، بعد ویژگی‌های GP و NV-DIS اضافه می‌شوند، درنهایت ویژگی‌های PMI و CSA در لایه سوم استفاده می‌شوند. همچنین شبکه‌ای با دولایه پنهان با ویژگی‌های GP، NV-DIS و CSA پیشنهادشده است. برای آزمایش‌ها از پایگاه داده‌های AVEC و IEMOCAP استفاده می‌شود. با همجوشی سلسله‌مراتبی نتایج بهتری نسبت به همجوشی سطح ویژگی و سطح تصمیم‌گیری به دست می‌آید [۱۳].

کایا[9] و همکاران، RNN-LSTM را برای تشخیص هیجانات آکوستیک corpus-cross و وظیفه متقاطع با استفاده از مدل عاطفی بررسی کردند. آن‌ها از رویکردی برای استفاده از ظرفیت سطح فریم و پیش‌بینی‌های برانگیختگی مدل‌های LSTM برای طبقه‌بندی هیجانات سطح بیان استفاده کردند و پیش‌بینی‌های گسسته مدل‌های LSTM را با اجزای سیستم پایه ترکیب کردند. سیستم پایه استفاده‌شده از SVM و عملکرد یادگیرنده با استفاده از هسته وزنی مبتنی بر حداقل مربعات افزایش

[1] Peng.
[2] Zhang.
[3] Trigeorgis.
[4] Lim.
[5] Wu.
[6] Meta-Decision tree.
[7] Eyben.
[8] Tian.
[9] Kaya.

بیشتری یافته است. نتایج حاصل از LSTM و سیستم پایه با سطح امتیاز همجوشی وزنی ترکیب‌شده است. برای آزمایش corpus-cross از مجموعه داده‌های RECOLA، SEMAINE و CreativeIT استفاده‌شده است. نتایج آن‌ها مناسب بودن روش پیشنهادی را برای امر تشخیص هیجان آکوستیک بین-پیکره در سطح زمانی پیوسته و گفتاری نشان داد [۱۴].

هوانگ[1] و همکاران یک شبکه عصبی برگشتی پیچشی عمیق با مکانیسم توجه را پیشنهاد کردند. آن‌ها نقش CNN را برای تشخیص هیجانات گفتار بررسی کردند، هم‌بستگی طیفی CNN را با تبدیل کسینوس گسسته تحت شرایط تمیز و نویز دار مقایسه کردند و اطلاعات زمینه را برای تولید وزن توجه بررسی کردند. سیستم پیشنهادی شامل یک‌لایه کانولوشن، یک‌لایه زمانی مبتنی بر BLSTM، یک‌لایه توجه کانولوشن و یک‌لایه کاملاً متصل است. لایه کانولوشن، بازنمایی سطح بالا را از Mels-log استخراج می‌کند که اطلاعات تکمیلی را در مقایسه با سیگنال شکل موج خام ارائه می‌دهد و همچنین اجازه می‌دهد تا به‌طور مستقیم مزیت هم‌بستگی مختص CNN را نسبت به DCT تعیین کند. سپس بازنمایی سطح بالا استخراج‌شده برای یادگیری وابستگی‌های زمانی به BLSTM وارد می‌شود. پس‌ازآن، لایه توجه کانولوشنی به‌صورت محلی اطلاعات زمینه و وزن‌ها را جمع‌آوری می‌کند. درنهایت، خروجی لایه توجه برای طبقه‌بندی به یک‌لایه کاملاً متصل وارد می‌شود. روش پیشنهادی با یک SVM پایه مقایسه شده است. مجموعه داده eNTERFACE' برای آزمایش انتخاب‌شده است. دقت روش پیشنهادی در شرایط نویز دار ۸۴٫۰۰٪ است [۱۵].

در سال ۲۰۲۱، هان[2] و لنگ[3]، برای تشخیص هیجان گفتار یک شبکه موازی Transformer-CNN-ResNet پیشنهاد دادند. آن‌ها به‌منظور کاهش مشکلات ناشی از عمیق شدن شبکه از ResNet استفاده کردند. استدلال آن‌ها این بود که CNN پارامترهای کمتری را برای افزایش توانایی بیان برازش شبکه محاسبه می‌کند. آن‌ها از شبکه عصبی بازگشتی سنتی، با وابستگی بلندمدت به استخراج ویژگی دنباله‌های گفتار و متن پرداختند، همچنین از مکانیسم توجه برای موازی کردن دنباله استفاده کردند. آن‌ها ارزیابی را بر روی مجموعه داده RAVDESS انجام دادند که دقت روش آن‌ها برای ۸ هیجان ۸۰٫۸۹ به دست آمد [۱۶].

چن[4] و ژینگ[5] در سال ۲۰۲۳، از یک ترنسفورمر تغییرپذیر مبتنی بر پنجره به‌منظور تشخیص هیجان گفتار استفاده کردند. استدلال آن‌ها این بود که در مقایسه با مکانیسم اصلی توجه کامل، توجه مبتنی بر پنجره در یادگیری ویژگی‌های دانه‌ریز[6] مؤثرتر است و درعین‌حال افزونگی مدل را تا حد زیادی کاهش می‌دهد. بااین‌حال، نشانه‌های احساسی به شیوه‌ای چنددانه‌ای وجود دارد، به‌طوری‌که پنجره ثابت از پیش تعریف‌شده می‌تواند انعطاف‌پذیری مدل را به‌شدت کاهش دهد. آن‌ها مدل را بر روی دیتاست IEMOCAP ارزیابی کردند که به‌دقت ۷۱٫۸ رسیدند [۱۷].

آندایانی[7] و تنگ[8] یک سیستم ترکیبی شامل LSTM و ترنسفورمر را به‌منظور تشخیص هیجان گفتار در سال ۲۰۲۳ پیشنهاد دادند. استدلال آن‌ها این بود که در هرکدام از آن‌ها محدودیت‌هایی وجود دارد. درنتیجه، آن‌ها روش ترکیبی را برای یادگیری وابستگی‌های بلندمدت در سیگنال‌های گفتار و طبقه‌بندی احساسات موردبحث استفاده کردند. آن‌ها ویژگی MFCC را استخراج کردند و به طبقه‌بندی کننده Transformer-LSTM کردند. دقت مدل آن‌ها با مجموعه داده‌های RAVDESS، DB-Emo به ترتیب ۷۵٫۶۲٪، ۸۵٫۵۵٪ رسید [۱۸].

در سال ۲۰۲۲ کومار[9] و مهارانا[10]، مدل ترکیبی LSTM-CNN و ترنسفورمر را برای تشخیص هیجان گفتار مورداستفاده قراردادند. مدل LSTM-CNN مورداستفاده برای تشخیص احساسات توسط چهار لایه CNN، یک لایه کاملاً متصل LSTM است. CNN شامل لایه کانولوشن یک‌بعدی، لایه نرمال‌سازی دسته‌ای، لایه تابع فعال‌سازی و لایه Pooling Max است. آن‌ها ابتدا ویژگی MFCC را از دیتاست DB-EMO شامل چهار هیجان غم، شادی، عصبانی و خنثی استخراج کردند، سپس مدل خود را به‌صورت جداگانه بر روی مدل CNN-LSTM و ترنسفورمر ارزیابی کردند، در مدل LSTM-CNN به‌دقت ۸۵٫۳ و در مدل ترنسفورمر به‌دقت ۸۸٫۵ رسیدند [۱۹].

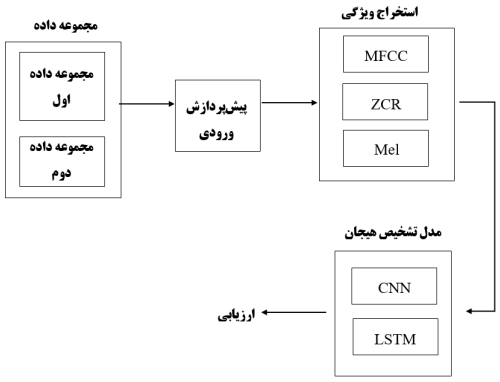
به‌منظور تشخیص هیجان گفتار، دات و گادر[11] در سال ۲۰۲۳ به بررسی تأثیر تبدیل موجک در سیستم‌های تشخیص هیجان در مقایسه با تبدیل فوریه پرداختند. آن‌ها عقیده داشتند که تبدیل فوریه به دلیل اصل عدم قطعیت، یک محدودیت اساسی دارد که به‌طور هم‌زمان امکان وضوح خوب در هر دو حوزه زمان و فرکانس را نمی‌دهد. به همین دلیل یک روش تشخیص عمیق هیجان گفتار مبتنی بر موجک را با استفاده از رمزگذار خودکار و شبکه‌های عصبی کانولوشنال ۱ بعدی و شبکه‌های حافظه کوتاه‌مدت پیشنهاد دادند. رمزگذار خودکار را برای انجام کاهش ابعاد ویژگی‌های موجک استفاده کردند، سپس از فضای پنهان[12] برای طبقه‌بندی

[1] Huang.
[2] Han.
[3] Leng.
[4] Chen.
[5] Xing.
[6] fine-grained.
[7] Andayani.
[8] Theng.
[9] Kumar.
[10] Maharana.
[11] Dutt and Gader.
[12] latent space.

احساسات با استفاده از مدل ۱ بعدی LSTM-CNN استفاده کردند. آن‌ها به‌دقت ۸۱,۴۵ بر روی دیتاست RAVDESS که شامل ۸ هیجان است، رسیدند [۲۰].

۲-۲. روش کار

شکل (۱)، نمایانگر روش کلی کار تشخیص هیجان در گفتار با استفاده از CNN و LSTM است.

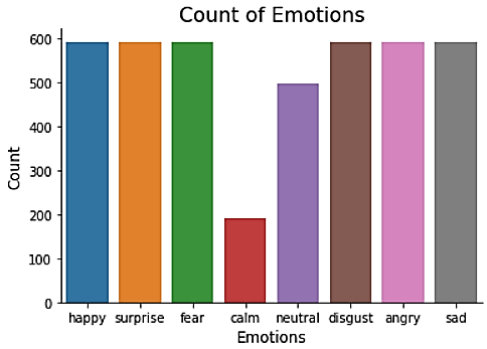


شکل (۱). روش کلی پژوهش

۲-۲-۱. مجموعه داده

در این پژوهش، به‌منظور داشتن تعداد داده‌های بیشتر، از ترکیب دو مجموعه داده موجود، یک مجموعه داده کلی جمع‌آوری شد. اولین مجموعه داده مورداستفاده، مجموعه داده RAVDESS است. این پایگاه داده شامل ویدئوها و ضبط‌های صوتی از گفتارها و صداهای هیجانی به زبان انگلیسی است و به هشت هیجان مختلف، ازجمله خنثی، آرام، شاد، غم، خشم، ترس، نفرت و شگفت‌زده، دسته‌بندی‌شده است. این پایگاه داده شامل اطلاعات ۲۴ بازیگر حرفه‌ای است و هر بازیگر دارای ۶۰ آیتم سمعی و بصری و ۴۴ آیتم صوتی است که درمجموع ۱۰۴ آیتم می‌شود. هر تولید ضبط‌شده از یک بازیگر در سه فرمت مختلف موجود است: سمعی-بصری، فقط ویدئو و فقط صوت. هر بازیگر برای هر هیجان ۲ جمله اجرا می‌کند: "بچه‌ها کنار در صحبت می‌کنند" و "سگ‌ها کنار در نشسته‌اند". این عبارات نیز در دو شدت هیجانی متفاوت، عادی و قوی، برای هر هیجان، به‌جز خنثی (فقط عادی) ثبت شد. بازیگران هر جمله را دو بار تکرار کردند. مجموعه ۷۳۵۶ ضبط‌شده هرکدام ۱۰ بار ازنظر اعتبار عاطفی، شدت و صحت رتبه‌بندی شدند. اعتبارسنجی RAVDESS با ۲۴۷ ارزیاب از آمریکای شمالی انجام شد. اعتبار به دقتی اشاره دارد که شرکت‌کنندگان با آن هیجانات موردنظر بازیگران را به‌درستی شناسایی کردند. نمرات بالابود و ۸۰ درصد برای صوتی-تصویری، ۷۵ درصد فقط برای ویدئو و ۶۰ درصد برای فقط صوتی به دست آورد [۲۱].

مجموعه داده دوم به نام TESS است. در این پایگاه داده، مجموعه‌ای از ۲۰۰ کلمه هدف در عبارت "کلمه _____ را بگویید" توسط دو بازیگر زن (۲۶ و ۶۴ ساله) گفته شد و ضبط‌هایی از مجموعه انجام شد که هر یک از هفت هیجان (خشم، نفرت، ترس، شادی، خوشایند، شگفت‌زده، غم و خنثی) را به تصویر می‌کشد. درمجموع ۲۸۰۰ فایل ضبطی وجود دارد. دو بازیگر زن از منطقه تورنتو استخدام شدند. هر دو بازیگر زن به‌عنوان زبان‌اصلی خود انگلیسی صحبت می‌کنند، تحصیلات دانشگاهی دارند و آموزش موسیقی دیده‌اند، فرمت فایل صوتی فرمت WAV است [۲۲]. در شکل(۲)، نیز تعداد داده‌های موجود در هر یک از کلاس‌های دیتاست ترکیب‌شده را می‌توان مشاهده کرد.



شکل (۲). تعداد داده‌های موجود در هر یک کلاس هیجان

۲-۲-۲. پیش‌پردازش ورودی

در مرحله پیش‌پردازش داده‌ها، هدف اصلی این است که داده‌های مجموعه داده را از اثرات مخربی که ممکن است خطاها را افزایش دهند و دقت مدل را کاهش بدهند، پاک‌سازی کرده یا اصلاحات لازم را بر روی آن‌ها اعمال کرد. برای این منظور، تعدادی فن پیش‌پردازش ازجمله Noise، Stretch و Shift به کار گرفته می‌شود. فن Noise باهدف حذف سروصداهای مختلفی که ممکن است در سیگنال‌های گفتاری ناشی از منابع مختلف وجود داشته باشند، به کار می‌رود. این نویزها معمولاً ویژگی‌های سیگنال‌های گفتار را تغییر می‌دهند و ممکن است کیفیت و درک گفتار را کاهش دهند. فن Stretch برای تغییر سرعت یا مدت‌زمان سیگنال صوتی بدون تأثیر منفی بر سرعت آن به کار می‌رود. این تغییرات معمولاً به‌منظور تطابق بهتر سیگنال‌های ورودی با مدل‌های گفتاری انجام می‌شود. همچنین، فن Shift برای تغییر از یک سبک گفتار به سبک دیگر استفاده می‌شود. این تغییرات به‌منظور ایجاد تنوع در داده‌های ورودی مدل و افزایش توانایی مدل در تعریف و تفکیک انواع مختلف گفتار اعمال می‌شوند. این اقدامات پیش‌پردازش می‌توانند به بهبود کارایی مدل‌های گفتاری و تشخیص هیجان در داده‌های ورودی کمک کنند.

۲-۲-۳. استخراج ویژگی

درروش پژوهش اجراشده برای بخش استخراج ویژگی، به‌جای انتخاب یک رویکرد مشخص برای استخراج ویژگی، از اجماع سه رویکرد مختلف برای استخراج ویژگی‌ها استفاده‌شده است. این نوآوری در این

بخش از پژوهش نشان‌دهنده تنوع در انتخاب ویژگی‌ها و بهبود عملکرد مدل است. سه رویکرد مورداستفاده برای استخراج ویژگی‌ها شامل MFCC (ضریب‌های آمپلی تود گذاری سیگنال‌های صوتی) نمایان در شکل (۳)، ZCR (نرخ چرخه‌های صفر) رابطه ریاضی (۱) و MelSpectrogram (طیف‌نمایی ملودیک) رابطه ریاضی (۲) هستند.

Algorithm 1: MFCC feature extraction algorithm pseudocode **input**: signal (Phonocardiogram signal) **output**: MFCC (MFCC of phonocardiogram signal)

function MFCC (parameters) Initialize parameters;

Split into frames phonocardiogram signals; Apply **Hamming windowing** to frames; Get spectrum by applying **Fast fourier transform** to all frames; Determine matrix for a mel-spaced filterbank, Transform spectrum to **mel spectrum**, Obtain **MFCC** vector for each frame by applying **discrete cosine transform** end function

شکل (۳). الگوریتم ضرایب‌های آمپلی تود گذاری سیگنال‌های صوتی

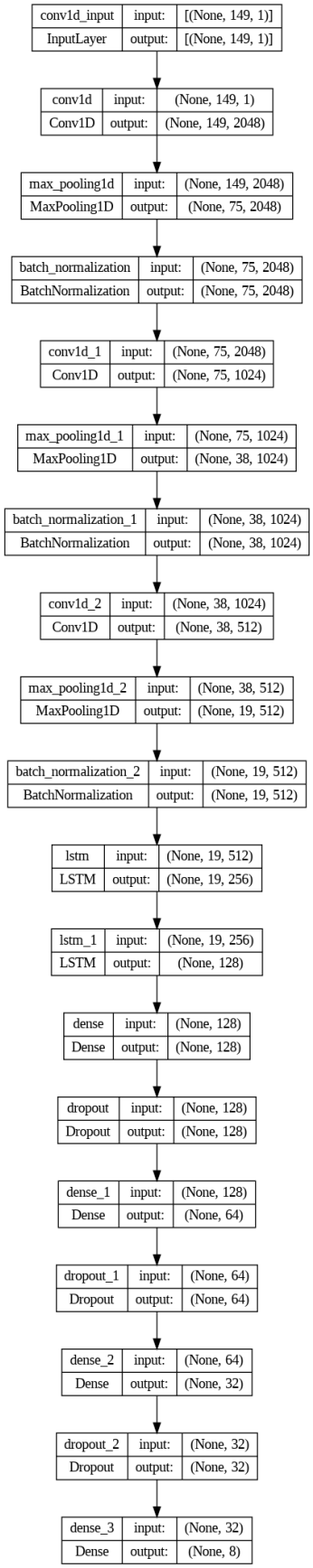
$$ZCR=1/N-1\sum_{n=1}^{N-1}|sgn(s[n])-sgn(s[n-1])| \quad (۱)$$

$$Mel\,(f)=2595\cdot\log_{10}(1+700f) \quad (۲)$$

فرایند استخراج ویژگی‌ها بر روی تمامی سیگنال‌های ورودی انجام می‌شود. بعد از استخراج ویژگی‌ها برای هر داده، سه مجموعه ویژگی به‌دست‌آمده از این سه رویکرد موردبررسی قرار می‌گیرند. در این مرحله، به‌منظور انتخاب بهترین مجموعه ویژگی‌ها از هر سه رویکرد، یک بررسی دقیق انجام می‌شود. این بررسی شامل مقایسه و تحلیل ویژگی‌ها و ارتباط آن‌ها با اهداف مدل‌سازی می‌شود. نتیجه این بررسی به انتخاب بهترین مجموعه ویژگی‌ها از هر سه رویکرد منجر می‌شود. این مجموعه بهترین ویژگی‌ها سپس ذخیره می‌شوند و برای مراحل بعدی تحلیل و پردازش به کار گرفته می‌شوند. این رویکرد متناسب با اهداف پژوهش و مطالب مقاله، بهبود مؤثری در عملکرد مدل‌های موردمطالعه فراهم می‌کند.

۴-۲-۲. مدل تشخیص هیجان

در این بخش، مدل تشخیص هیجان به‌صورت دقیق توضیح داده‌شده و نحوه استفاده از شبکه‌های عصبی کانوولوشنال و شبکه‌های عصبی با حافظه کوتاه‌مدت بلند برای تشخیص هیجان از صوت و گفتار توضیح داده می‌شود. مدل تشخیص هیجان از صوت و گفتار با استفاده از یک ترکیب از شبکه‌های عصبی کانوولوشنال (CNN) و شبکه‌های عصبی با حافظه کوتاه‌مدت بلند (LSTM) طراحی و پیاده‌سازی شده است. این مدل به‌منظور تشخیص هیجان در داده‌های صوتی باتوجه‌به نوع داده‌های ورودی و ویژگی‌های مرتبط با هیجان، تنظیم می‌شود.



شکل (۴). ساختار کلی مدل تشخیص هیجان

ساختار مدل به‌صورت یک توالی از لایه‌ها تعریف می‌شود و شامل مراحل زیر است:

- لایه ورودی: در ابتدا، یک‌لایه ورودی به مدل اضافه می‌شود. این لایه ورودی برای دریافت داده‌های ورودی تعریف می‌شود.
- مدل CNN: مدل شبکه عصبی کانولوشنال دارای سه لایه Convolutional با هسته‌های مختلف است. این لایه‌ها برای استخراج ویژگی‌های صوتی از داده‌های ورودی به کار می‌روند. بین این لایه‌ها، یک‌لایه Normalization Batch برای نرمال‌سازی داده‌ها و بهبود عملکرد شبکه اضافه‌شده است.
- مدل LSTM: خروجی مدل CNN به‌عنوان ورودی به مدل LSTM منتقل می‌شود. مدل LSTM شامل دولایه است و برای تحلیل توالی‌های زمانی و ویژگی‌های استخراج‌شده از مدل CNN به کار می‌روند.
- لایه‌های Dense و Dropout: سه لایه Dense به مدل اضافه‌شده‌اند. پس از هر لایه Dense، سه لایه Dropout استفاده می‌شود تا از افزایش وزن یادگیری شده در شبکه جلوگیری کند و از بیش برازش در مدل جلوگیری شود.
- لایه خروجی: درنهایت، یک‌لایه Dense نهایی با ۸ کلاس خروجی تعریف می‌شود که متناظر با هشت احساس مختلف مورد تشخیص قرار می‌گیرد.

شبکه‌های عصبی کانولوشنال (CNN) به‌خوبی برای استخراج ویژگی‌های صوتی از داده‌های ورودی به کار می‌روند. در مدل پیشنهادی، این لایه‌ها برای استخراج ویژگی‌های صوتی مرتبط با هیجان از داده‌های صوتی استفاده می‌شوند. ازآنجاکه داده‌ها ممکن است صوت‌هایی از گفتار افراد را شامل شوند، CNN می‌تواند اطلاعات مرتبط با احساسات از گفتار استخراج کند.

با داده‌های ویژگی‌های صوتی به‌دست‌آمده از مدل CNN، مدل LSTM به تحلیل توالی‌های زمانی و ویژگی‌های استخراج‌شده می‌پردازد. LSTM به‌عنوان یک شبکه با حافظه کوتاه‌مدت بلند مناسب برای مدل‌سازی داده‌های دنباله‌ای و توالی‌های زمانی است. با استفاده از این ترکیب از CNN و LSTM، مدل تشخیص هیجان قادر به تشخیص و تمییز هیجانات مختلف از داده‌های صوتی می‌شود.

۳. نتایج و بحث

در این بخش، ابتدا به بررسی کلی رویکردهای مشابه با حوزه حال می‌پردازیم. سپس با تمرکز بر روی رویکردهای شبیه‌سازی، به بررسی دقیق‌تر این روش‌ها در مقایسه با روش پیشنهادی خودخواهیم پرداخت. نتایج به‌دست‌آمده از ارزیابی مدل پیشنهادی برای تشخیص هیجان ارائه می‌شوند. باتوجه‌به رابطه (۳) که دقت مدل را نشان می‌دهد، مشخص می‌شود که مدل پیشنهادی توانست دقت ۹۱ درصد را دست‌یابی کند.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (۳)$$

$$Precision = \frac{TP}{TP + FP} \quad (۴)$$

$$Recall = \frac{TP}{TP + FN} \quad (۵)$$

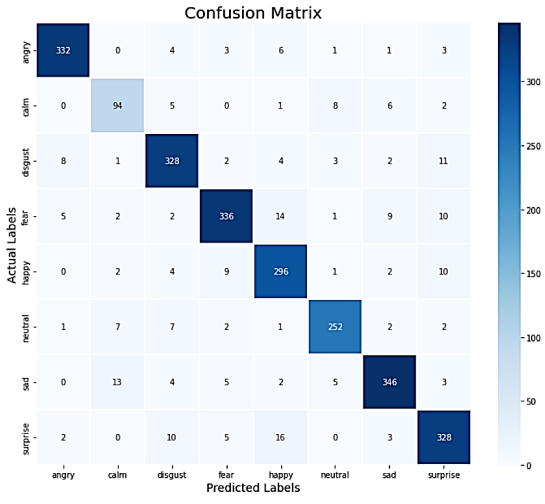
$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (۶)$$

TN = negative True ,FN = negative False ,TP = positive True ,FP = positive False

بعد از محاسبه دقت مدل پیشنهادی بر اساس روابط (۳) تا (۶)، معیارهای Precision، Recall و F۱-score برای تمامی کلاس‌های پژوهشی محاسبه‌شده است. نتایج به‌دست‌آمده که در جدول (۱) نمایش داده‌شده است، نشان می‌دهد که مدل پیشنهادی توانست عملکرد بسیار مؤثری در تمامی کلاس‌های پژوهشی داشته باشد. همچنین، با توجه به شکل (۵)، ماتریس درهم‌ریختگی برای هر کلاس محاسبه‌شده و نشان می‌دهد که مدل توانسته است هیجانات مختلف را با دقت و صحت بالایی تشخیص دهد و مرتبط با آن‌ها عمل کند. به‌علاوه، نرخ تشخیص مدل پیشنهادی نیز موردبررسی قرار گرفت. نرخ تشخیص نشان‌دهنده توانایی مدل در تشخیص مثبت‌ها (صحیح مثبت) از مجموع تمامی موارد مثبت است. در مورد تشخیص هیجان از گفتار، این نرخ نشان می‌دهد که چقدر مدل توانسته است هیجانات موردنظر را به‌درستی شناسایی کرده و تشخیص دهد. نتایج نرخ تشخص نیز نشان می‌دهند که مدل پیشنهادی توانایی بالایی در تشخص هیجانات از گفتار دارد و عملکرد عالی در این زمینه ارائه می‌دهد. این نرخ تشخص نقش مهمی در ارزیابی عملکرد مدل و کارایی آن در تشخص هیجان از گفتار ایفا می‌کند. این مدل با دقت بالا و توانایی در تعبیر از هیجانات انسانی در داده‌های صوتی و ویژگی‌های صوتی، توانسته است در تشخیص و شناسایی انواع هیجانات مختلف، ازجمله خنثی، آرام، شاد، غم، خشم، ترس، نفرت و شگفت‌زده، به‌صورت موفقیت‌آمیز عمل کند. این مدل قابلیت اعتمادی را در شناسایی هیجانات از داده‌های صوتی ارائه می‌دهد و می‌تواند به مسائل و برنامه‌های متعددی که به تشخیص هیجانات از صوت و گفتار نیاز دارند، اعمال شود.

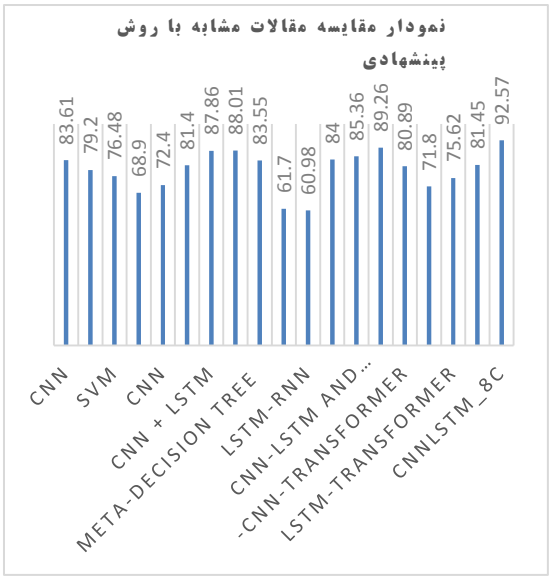
جدول (۱). نتایج کلی ارزیابی مدل و کلاس‌ها

هیجان	precision	recall	score-۱f	support
خشم	۰,۹۶	۰,۹۵	۰,۹۵	۳۵۰
آرام	۰,۸۳	۰,۷۸	۰,۸۱	۱۱۶
نفرت	۰,۸۹	۰,۹۴	۰,۹۱	۳۵۹
ترس	۰,۹۵	۰,۹۲	۰,۹۳	۳۷۹
شادی	۰,۹۱	۰,۹۱	۰,۹۱	۳۲۴
خنثی	۰,۹۲	۰,۹۳	۰,۹۳	۲۷۴
غم	۰,۹۱	۰,۹۳	۰,۹۲	۳۷۸
شگفت‌زده	۰,۹۴	۰,۹۳	۰,۹۳	۳۶۴
accuracy		۰,۹۲		۲۵۴۴
macro avg	۰,۹۱	۰,۹۱	۰,۹۱	۲۵۴۴
weighted avg	۰,۹۲	۰,۹۲	۰,۹۲	۲۵۴۴

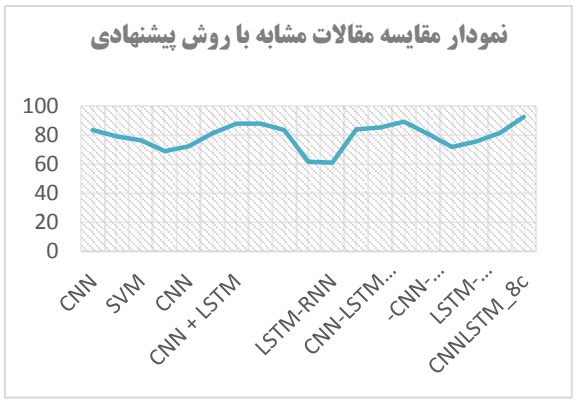


شکل (۵). ماتریس درهم‌ریختگی

یکی از روش‌های پیشنهادی برای این حوزه، استفاده از مدل‌های یادگیری عمیق مانند شبکه عصبی کانولوشنالی است که به کمک آن‌ها می‌توان با دقت بالا به تشخیص هیجان از گفتار پرداخت. در اینجا، ما به تحلیل مقالات مرتبط با این حوزه و مقایسه آن‌ها با روش پیشنهادی می‌پردازیم. جدول(۲)، نتایج تحلیلی مقالات مشابه را نشان می‌دهد که از روش‌های مختلفی برای تشخیص هیجان استفاده کرده‌اند. در مقایسه با مقالات مشابه، روش پیشنهادی که از ترکیب شبکه عصبی کانولوشنال و شبکه‌های عصبی با حافظه کوتاه‌مدت بلند برای تشخیص هیجان استفاده می‌کند، نتایج بهتری را به دست آورده است. این روش با دقت بالا و حساسیت مناسب، قادر به استخراج ویژگی‌های پربار از گفتار است که به تشخیص دقیق‌تر هیجانات کمک می‌کند. به‌طورکلی، این مقایسه نشان می‌دهد که استفاده از روش پیشنهادی می‌تواند بهبود قابل‌توجهی در عملکرد تشخیص هیجان از گفتار داشته باشد. اشکال (۶) و (۷) نیز می‌تواند نمایشی از نتایج به‌دست‌آمده باشد:



شکل (۶). نمودار مقایسه روش پیشنهادی با روش‌های دیگر



شکل (۷). نمودار مقایسه روش پیشنهادی با روش‌های دیگر

باتوجه‌به نتایج مطرح‌شده در جدول و شکل‌ها، می‌توانیم نتیجه بگیریم که روش پیشنهادی ارائه‌شده، عملکرد بهتری در تشخیص هیجان از گفتار ارائه می‌دهد نسبت به روش‌های مشابه موجود در این حوزه.

جدول (۲). ارزیابی مدل‌های مشابه تشخیص هیجان از گفتار

ردیف	مدل	دقت	مرجع
۱	CNN	۸۳,۶۱	[۵]
۲	RNN	۷۹,۲۰	
۳	SVM	۷۶,۴۸	
۴	MVR	۶۸,۹۰	
۵	CNN	۷۲,۴	[۶]
۶	CNN	۸۱,۴	[۷]
۷	LSTM + CNN	۸۷,۸۶	[۸]
۸	CNN	۸۸,۰۱	[۱۰]
۹	tree Decision-Meta	۸۳,۵۵	[۱۱]
۱۰	LSTM	۶۱,۷	[۱۳]
۱۱	RNN-LSTM	۶۰,۹۸	[۱۴]
۱۲	LSTM + CNN	۸۴,۰۰	[۱۵]
۱۳	and LSTM-CNN Transformer Vision	۸۵,۳۶	[۱۹]
۱۴	LSTM + CNN	۸۹,۲۶	[۲۳]
۱۵	Transformer-CNN-	۸۰,۸۹	[۱۶]
۱۶	Speech Deformable Transformer	۷۱,۸	[۱۷]
۱۷	Transformer-LSTM	۷۵,۶۲	[۱۸]
۱۸	Wavelet TM, gLS CNN	۸۱,۴۵	[۲۰]
۱۹	LSTM + CNN	۹۲,۵۷	روش پیشنهادی

در این بخش، پس از مرور کلی رویکردهای مشابه در حوزه، به تحلیل دقیق‌تر شبیه‌سازی رویکردهای مشابه با روش پیشنهادی می‌پردازیم.

جدول (۳). ارزیابی مدل‌های مشابه تشخیص هیجان از گفتار از منظر شبیه‌سازی

مقالات	مدل	تعداد لایه ها	استخراج کننده ویژگی	دقت
[۵]	CNN	۸	Mel-frequency cepstral coefficients (MFCCs)	۸۳٫۶۱
[۱۵]	CNN + LSTM	۱۰	MFCCs	۸۴
[۲۳]	CNN + LSTM		Mel Spectrogram-Vision Transformer	۸۹٫۲۶
[۲۴]	CNN-14 of the PANNs framework + Spatial Transformer Network on saliency maps and facial images followed by a bi-LSTM with an attention mechanism	۱۴	MFCCs with spectral centroids	۸۰.۸٪
[۲۵]	CNN and LSTM networks			۷۱/۱۸
[۲۶]	Deep CNN and RESNET	۲۵	MFCC, MFC, and Chroma	۶۵.۹۷٪
[۲۷]	deep convolutional neural networks	۱۲	MFCCs, chromagram, mel-scale spectrogram, Tonnetz representation, and spectral contrast features	۷۱/۶۱
[۲۸]	VGG-16 + BiLSTM	۱۶	MFCCs	۸۷/۳
رویکرد پیشنهادی	Hybrid CNN-LSTM	۱۹	Mel-frequency cepstral coefficients (MFCCs),MEL , ZCR	۹۲/۵۷

رویکرد پیشنهادی در مقایسه با ۹ مقاله دیگر عملکرد بسیار خوبی دارد و ازنظر دقت، بالاترین رتبه را به خود اختصاص داده است. این نشان می‌دهد که مدل پیشنهادی در تشخیص هیجان از گفتار بسیار کارآمد است.

۴. نتیجه‌گیری

در این مقاله، مدلی ترکیبی مبتنی بر CNN و LSTM و بر پایه سه ویژگی MFCC، MEL و ZCR جهت تشخیص هشت هیجان پایه سیگنال گفتار ارائه شد، نتایج ارزیابی مدل، دقت متوسط ۹۲,۵۷ درصد برای هیجان غم، شادی، خشم، نفرت، آرام، شگفت شده، خنثی و ترس را نشان می‌دهد، دقت مدل در هیجان خشم ۹۶ درصد، ترس ۹۵ درصد و در هیجان شگفت‌زده، ۹۴ درصد است، همچنین نتیجه ارزیابی مدل‌های ارائه‌شده در جدول(۱)، با مدل ارائه‌شده در این پژوهش، عملکرد بهتر این مدل را ۳,۳۱ درصد بهبود نسبت به بهترین مدل ارائه‌شده در این جدول را نشان می‌دهد. مهم‌ترین دلایل برتری مدل پیشنهادی نسبت به مقالات دیگر، استفاده از ویژگی‌های مناسب‌تر و تعداد لایه¬های مناسب است.

۵. مراجع

[1] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review, " International Journal of Speech Technology, vol. 21, pp. 93-120, 2018. DOI:https://doi.org/10.1007/s10772-018-9491-z

[2] M. Shamsi, "Emotion Recognition in Persian Speech Using Machine Learning Methods," 2016. DOI:10.1109/ICCKE54056.2021.9721504

[3] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, " Speech Communication, vol. 116, pp. 56-76, 2020.. DOI: https://doi.org/10.1016/j.specom.2019.12.001

[4] B. Du, Q. Gao, and H. Ning, "Survey on Intelligent Speech Emotion Recognition, " Forest Chemicals Review, pp. 230-260, 2021. DOI: 10.1109/SLT.2016.7846319

[5] A. B. A. Qayyum, A. Arefeen, and C. Shahnaz, "Convolutional neural network (CNN) based speech-emotion recognition, " in 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), 2019, pp. 122-125. DOI:10.1109/SPICSCON48833.2019.9065172

[6] F. Makhmudov, A. Kutlimuratov, F. Akhmedov, M. S. Abdallah, and Y.-I. Cho, "Modeling Speech Emotion Recognition via Attention-Oriented Parallel CNN Encoders," Electronics, vol. 11, p. 4047, 2022. DOI: https://doi.org/10.3390/electronics11234047

[7] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale cnn and attention, " in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 3020-3024. DOI: 10.1109/ICASSP39728.2021.9414286

[8] H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu, and G. Dai, "Pre-trained deep convolution neural network model with attention for speech emotion recognition, " Frontiers in Physiology, vol. 12, p. 643202, 2021. DOI:10.3389/fphys.2021.643202

[9] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, " in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2016, pp. 5200-5204. DOI: 10.1109/ICASSP.2016.7472669

[10] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks, " in 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), 2016, pp. 1- DOI: 10.1109/APSIPA.2016.7820699

[11] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," IEEE Transactions on Affective Computing, vol. 2, pp. 10-21, 2010. DOI: 10.1109/T-AFFC.2010.16

[12] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues, " Journal on Multimodal User Interfaces, vol. 3, pp. 7-19, 2010 DOI: https://doi.org/10.1007/s12193-009-0032-6

[13] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features, " in 2016 IEEE Spoken Language Technology Workshop (SLT), 2016, pp. 565-572. DOI: 10.1109/SLT.2016.7846319

[14] H. Kaya, D. Fedotov, A. Yesilkanat, O. Verkholyak, Y. Zhang, and A. Karpov, "LSTM Based Cross-corpus and Cross-task Acoustic Emotion Recognition, " in Interspeech, 2018, pp. 521-525. DOI:10.21437/Interspeech.2018-2298

[15] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition, " in 2017 IEEE international conference on multimedia and expo (ICME), 2017, pp. 583-588. DOI: 10.1109/ICME.2017.8019296

[16] S. Han, F. Leng, and Z. Jin, "Speech emotion recognition with a ResNet-CNN-Transformer parallel neural network, " in 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), 2021, pp. 803-807. DOI: 10.1109/CISCE52179.2021.9445906

[17] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "DST: Deformable Speech Transformer for Emotion Recognition, " in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1-5. DOI: 10.1109/ICASSP49357.2023.10096966

[18] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-transformer model for emotion recognition from speech audio files, " IEEE Access, vol. 10, pp. 36018-36027, 2022. DOI: 10.1109/ACCESS.2022.3163856

[19] C. A. Kumar, A. D. Maharana, S. M. Krishnan, S. S. S.

Hanuma, G. J. Lal, and V. Ravi, "Speech Emotion Recognition Using CNN-LSTM and Vision Transformer, " in International Conference on Innovations in Bio-Inspired Computing and Applications, 2022, pp. 86-97. DOI: https://doi.org/10.1007/978-3-031-27499-2_8

[20] A. Dutt and P. Gader, "Wavelet Multiresolution Analysis Based Speech Emotion Recognition System Using 1D CNN LSTM Networks, " IEEE / ACM Transactions on Audio, Speech, and Language Processing, 2023. DOI: 10.1109/TASLP.2023.3277291

[21] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, " PloS one, vol. 13, p. e0196391, 2018. DOI: https://doi.org/10.1371/journal.pone.0196391. g001

[22] K. Dupuis and M. K. Pichora-Fuller, "Toronto emotional speech set (TESS) -Younger talker_Happy," 2010. DOI: https://doi.org/10.5683/SP2/E8H2MF

[23] B. Salian, O. Narvade, R. Tambewagh, and S. Bharne, "Speech Emotion Recognition using Time Distributed CNN and LSTM, " in ITM Web of Conferences, 2021, p. 03006. DOI: https://doi.org/10.1051/itmconf/20214003006

[24] Luna-Jiménez, C.; Griol, D.; Callejas, Z.; Kleinlein, R.; Montero, J.M.; Fernández-Martínez, F. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. Sensors 2021, 21, 7665. https://doi.org/10.3390/s21227665

[25] Tanberk, S., Tükel, D.B. (2022). Ensemble Learning with CNN–LSTM Combination for Speech Emotion Recognition. In: Bashir, A.K., Fortino, G., Khanna, A., Gupta, D. (eds) Proceedings of International Conference on Computing and Communication Networks. Lecture Notes in Networks and Systems, vol 394. Springer, Singapore. https://doi.org/10.1007/978-981-19-0604-6_5

[26] Lakshmi, K.L., Muthulakshmi, P., Nithya, A.A. et al. Recognition of emotions in speech using deep CNN and RESNET. Soft Comput (2023). https://doi.org/10.1007 / s00500-023-07969-5

[27] D. Issa, M. Fatih Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks, ” Biomed. Signal Process. Control, vol. 59, p. 101894, 2020, doi: 10.1016/j.bspc.2020.101894.

[28] Dangol, R., Alsadoon, A., Prasad, P.W.C. et al. Speech Emotion Recognition UsingConvolutional Neural Network and Long-Short TermMemory. Multimed Tools Appl 79, 32917–32934 (2020). https://doi.org/10.1007/s11042-020-09693-w