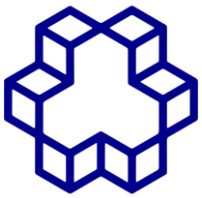
**به نام خدا**



دانشگاه صنعتی خواجه نصیرالدین طوسی

دانشکده برق

# مبانی مکاترونیک

## گزارش پروژه

## سیستم تشخیص ژست‌های دست

محمدمهدی کرمی　　احسان حبیبی　　علیرضا قربانی

**استاد : آقای دکتر مهدی دلربایی**

خرداد ماه ۱۴۰۴

# فهرست مطالب

عنوان شماره صفحه

## چکیده

هدف این پروژه طراحی و پیاده‌سازی یک سیستم شناسایی و ردیابی ژست‌های دست به کمک Mediapipe و یادگیری ماشین است. این سیستم قادر به شناسایی پنج ژست دست شامل " Thumbs Up"، "Thumbs Down"، "Left Swipe"، "Right Swipe" و "Stop" می‌باشد و از آن‌ها برای کنترل دستگاه‌های هوشمند به‌صورت بدون تماس فیزیکی استفاده می‌کند. ویژگی‌های استخراج‌شده از دست‌ها به‌عنوان ورودی به مدل یادگیری ماشین داده می‌شود تا به‌طور دقیق ژست‌ها را شناسایی کند و امکان تعامل بی‌دردسر با دستگاه‌ها را فراهم کند. این سیستم به‌ویژه برای کاربردهای خانگی نظیر کنترل تلویزیون‌های هوشمند، دستگاه‌های صوتی، و سیستم‌های تعاملی بدون نیاز به ریموت یا ورودی فیزیکی قابل استفاده است.

## مقدمه

با پیشرفت روزافزون فناوری‌های مختلف، تعاملات میان انسان و دستگاه‌ها به سمت سیستم‌های بدون تماس حرکت کرده است. این سیستم‌ها نه تنها راحتی و سرعت بیشتری به کاربران می‌دهند، بلکه در شرایطی که استفاده از ورودی‌های فیزیکی محدود است، می‌توانند مفید واقع شوند. یکی از این روش‌ها، استفاده از ژست‌های دست برای تعامل با دستگاه‌ها است. سیستم‌های شناسایی ژست دست، امکان کنترل دستگاه‌های مختلف نظیر تلویزیون‌های هوشمند، سیستم‌های صوتی، و حتی روبات‌ها را بدون نیاز به استفاده از ریموت کنترل یا صفحه‌کلید فراهم می‌کنند.

یکی از چالش‌های اصلی در شناسایی ژست‌های دست، شناسایی دقیق موقعیت دست و تشخیص صحیح ژست‌ها در شرایط محیطی مختلف است. روش‌هایی مانند Mediapipe که قادر به شناسایی ۲۱ لندمارک مختلف در دست هستند، به دلیل دقت بالا و سرعت پردازش مناسب، انتخاب خوبی برای این پروژه هستند. این سیستم به‌ویژه قادر است ژست‌های ایستا و حرکتی (مانند Left Swipe و Right Swipe) را با دقت بالا شناسایی کند و امکان تعامل بدون تماس را فراهم آورد.

### بیان مسئله

در بسیاری از سیستم‌های کنترلی، استفاده از ورودی‌های فیزیکی مانند دکمه‌ها یا صفحه‌کلید متداول است. با این حال، این روش‌ها محدودیت‌هایی دارند که شامل نیاز به فشردن دکمه‌ها و یا استفاده از صفحه‌کلید است. سیستم‌های شناسایی ژست‌های دست می‌توانند تجربه‌ای طبیعی‌تر و راحت‌تر را فراهم کنند. هدف این پروژه، طراحی و پیاده‌سازی یک سیستم شناسایی ژست دست است که از طریق

Mediapipe قادر به شناسایی دست‌ها و ژست‌های مختلف باشد و از آن‌ها برای تعامل با دستگاه‌های مختلف استفاده کند.

## پژوهش‌های پیشین

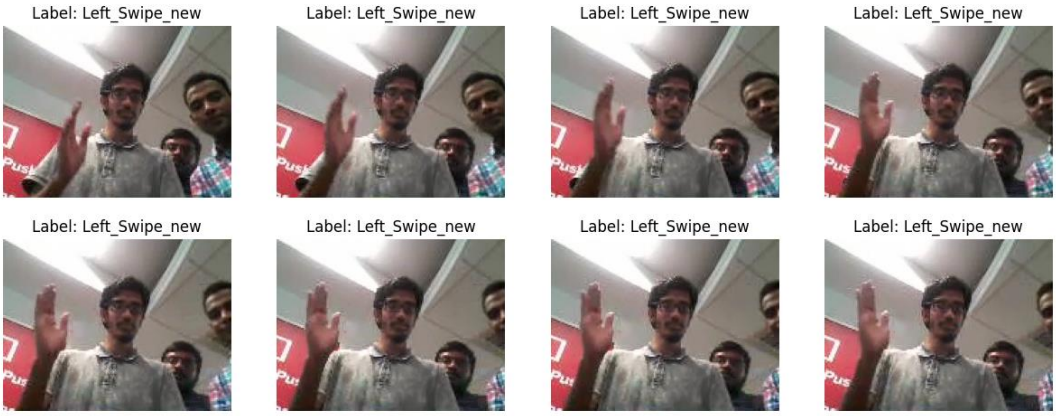
در مطالعات قبلی، الگوریتم‌های مختلفی برای شناسایی ژست‌های دست مورد استفاده قرار گرفته‌اند. یکی از پرکاربردترین و موفق‌ترین روش‌ها، استفاده از Mediapipe است. Mediapipe یک مدل پردازش تصویر است که برای شناسایی دست‌ها و صورت طراحی شده است و در پروژه‌های مختلفی برای شناسایی ژست‌های دست به‌ویژه در سیستم‌های هوشمند و بازی‌های ویدیویی به‌کار رفته است.

## رویکرد پیشنهادی

در این پروژه از Mediapipe برای شناسایی و ردیابی دست‌ها استفاده شده است. این سیستم قادر به شناسایی ۲۱ لندمارک مختلف در دست‌هاست که شامل نقاط مختلف انگشتان و مچ دست می‌شود. ویژگی‌هایی نظیر زاویه انگشت‌ها، فاصله بین انگشت‌ها و موقعیت مرکز دست از این نقاط استخراج شده و به مدل یادگیری ماشین داده می‌شود. مدل یادگیری ماشین سپس بر اساس این ویژگی‌ها آموزش دیده تا ژست‌های مختلف دست را شناسایی کند.

### داده‌های استفاده‌شده

برای این پروژه از مجموعه داده "Hand Gesture Detection System" که در پلتفرم Kaggle موجود است استفاده شد. این مجموعه داده شامل تصاویر متنوع از افراد مختلف است که پنج ژست مختلف دست را در شرایط نوری و محیطی مختلف به نمایش می‌گذارند. این داده‌ها به همراه برچسب‌های ژست‌های مربوطه برای آموزش و ارزیابی مدل یادگیری ماشین استفاده شدند.



### روش

**مرحله اول، آماده‌سازی داده‌ها:** ابتدا داده‌ها به دو بخش آموزش (train) و ارزیابی (val) تقسیم شدند. در این مرحله، لندمارک‌های دست با استفاده از Mediapipe از تصاویر استخراج شدند.

**مرحله دوم، استخراج ویژگی‌ها:** در این مرحله، برای استخراج ویژگی‌ها از پنجره‌ای متشکل از پنج فریم متوالی استفاده شد. برخلاف روش‌های ساده فریم به فریم که ممکن است باعث نوسان و خطا در پیش‌بینی شوند، این روش از پنجره‌ای به‌عنوان ورودی مدل استفاده می‌کند که ویژگی‌های ایستا (مانند زاویه انگشتان و فاصله‌ها) از فریم میانی و ویژگی‌های دینامیک (مانند حرکت دست) از تغییرات بین فریم اول و آخر محاسبه می‌شود.

**مرحله سوم، مدل یادگیری ماشین:** از یک مدل شبکه عصبی مصنوعی (ANN) با لایه‌های Dense برای آموزش مدل استفاده شد. مدل با ویژگی‌های استخراج‌شده از تصاویر آموزش دید و سپس برای پیش‌بینی ژست‌های جدید از آن استفاده شد.

**مرحله چهارم، ارزیابی مدل:** مدل با استفاده از داده‌های آموزش و ارزیابی مورد ارزیابی قرار گرفت و دقت مدل محاسبه شد.

## پیاده‌سازی

پس از آموزش مدل و به دست آوردن دقت بالا در داده‌های ارزیابی، مرحله بعدی تأیید عملکرد مدل در شرایط واقعی بود. برای این منظور، یک برنامه نوشته شد که با استفاده از وبکم لپ‌تاپ تصاویر دست را به صورت real-time دریافت می‌کند و همان پردازش‌های انجام شده روی داده‌های آموزش و ارزیابی، شامل استخراج ویژگی‌ها و پیش‌بینی ژست دست، روی این فریم‌های زنده اعمال می‌شود. این کار باعث شد تا مطمئن شویم که مدل تنها روی داده‌های آماده‌شده و از پیش پردازش شده عمل نمی‌کند و می‌تواند ژست‌ها را در شرایط واقعی، با نور و زاویه‌های مختلف شناسایی کند.

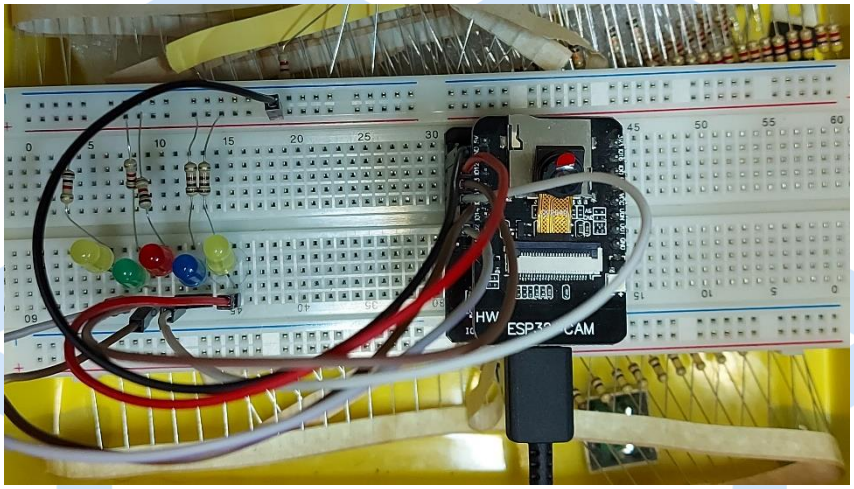
نتایج real-time نشان داد که مدل توانست ژست‌ها را با دقت قابل توجهی شناسایی کند و تغییرات سریع دست در طول پنجره ۵ فریمه، به درستی در ویژگی‌های دینامیک مدل لحاظ شد. این تجربه، اطمینان داد که مدل قابلیت پیاده‌سازی در محیط واقعی و نه صرفاً روی داده‌های آزمایشی را دارد.

### پیاده‌سازی عملی با ESP32-CAM

برای گام بعدی، یک پیاده‌سازی عملی با استفاده از ESP32-CAM انجام شد. این ماژول به صورت بی‌سیم تصاویر و ویدیوهای دست را به اکسس پوینت (Access Point) خود ارسال می‌کند. لپ‌تاپ به این

اکسس پوینت متصل شده و یک لینک ساده برای نمایش تصویر دوربین ESP ایجاد شد. با این کار، تصویر زنده دست‌ها در لپ‌تاپ قابل مشاهده بود و همان پردازش‌های real-time که پیش‌تر با وبکم لپ‌تاپ انجام شده بود، روی این تصویر اعمال شد.

ویژگی‌های استخراج‌شده از پنجره ۵ فریمه اعمال شد و پیش‌بینی ژست دست در زمان واقعی انجام گرفت. سپس نتایج پیش‌بینی از لپ‌تاپ دوباره به ESP32-CAM ارسال شد. بسته‌های داده شامل ژست تشخیص داده شده بودند و بر اساس آن، LEDهای متصل به ESP روشن یا خاموش می‌شدند تا کاربر بتواند نتیجه ژست خود را به صورت فیزیکی مشاهده کند.



این پیاده‌سازی نشان داد که مدل آموزش دیده می‌تواند بدون نیاز به اتصال مستقیم به کامپیوتر یا لپ‌تاپ، به صورت شبکه‌ای و با استفاده از ماژول ESP32-CAM کار کند و تعامل دست آزاد را به صورت بی‌سیم و real-time فراهم آورد. در عمل، این روش امکان توسعه سیستم برای کاربردهای خانگی مانند کنترل تلویزیون‌های هوشمند، دستگاه‌های صوتی یا سیستم‌های تعاملی بدون تماس را فراهم می‌کند.

## نتایج

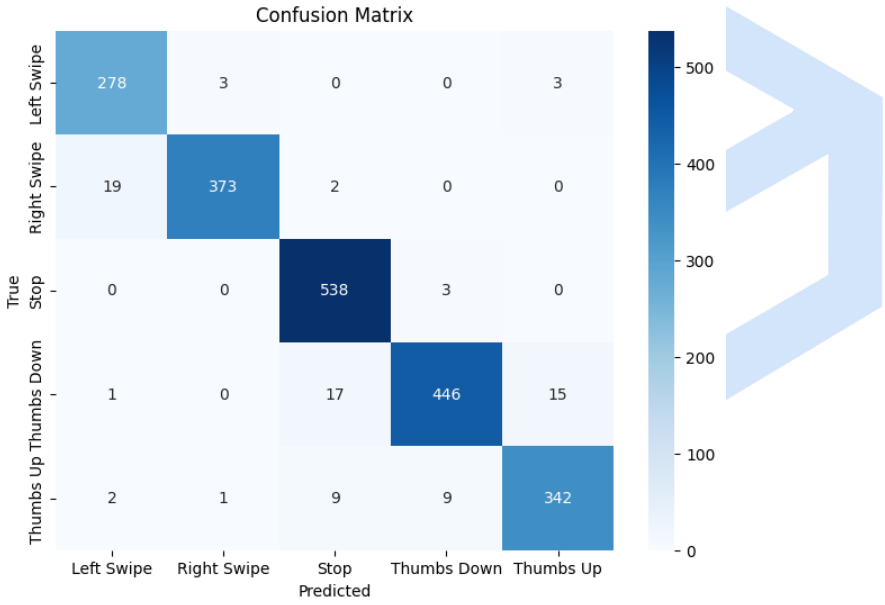
در این بخش، نتایج حاصل از آموزش مدل و ارزیابی آن بر روی داده‌های آزمایشی و ارزیابی جداگانه بررسی می‌شود. دو روش مختلف برای ارزیابی عملکرد مدل استفاده شده است: یکی بر اساس استفاده از پنجره‌های ۵ فریمه به صورت اسلایدینگ ویندو (Sliding Window) و مقایسه نتایج با برچسب‌های واقعی، و دیگری استفاده از رای‌گیری در سطح پوشه‌ها که برای داده‌های مربوط به هر حرکت (که به صورت مجموعه‌ای از فریم‌ها در یک پوشه ذخیره شده بودند) اعمال شد.

## نتایج آموزش مدل

مدل در ۳۰ اپوک آموزش داده شد و نتایج به‌دست آمده نشان داد که مدل به خوبی روی داده‌های آموزشی عمل کرده است. در آخرین اپوک، دقت مدل ٪۹۶.۲۲ و خطای آن ۰.۱۰۸۳ بود. این نشان‌دهنده عملکرد بسیار خوب مدل در شناسایی ژست‌ها در داده‌های آموزشی است. همچنین، در مجموعه داده‌های ارزیابی (Validation Data) نیز، دقت مدل به ٪۹۶.۲۱ رسید که نشان‌دهنده تعمیم خوب مدل به داده‌های جدید و دیده‌نشده بود.

## نتایج ارزیابی مدل با استفاده از Sliding Window

برای ارزیابی مدل، از پنجره‌های ۵ فریمه به صورت اسلایدینگ ویندو استفاده کردیم. در این روش، به ازای هر ۵ فریم متوالی، ویژگی‌ها استخراج شده و پیش‌بینی ژست دست انجام شد. نتایج حاصل از این روش ارزیابی به شرح زیر است:




مدل در داده‌های ارزیابی با دقت ٪۹۵.۹۲ عمل کرد. این دقت نشان می‌دهد که مدل توانسته است به خوبی ژست‌ها را در شرایط مختلف شناسایی کند.

```
Validation Accuracy: 0.9592

Classification Report:
              precision    recall  f1-score   support

  Left Swipe       0.93      0.98      0.95       284
 Right Swipe       0.99      0.95      0.97       394
        Stop       0.95      0.99      0.97       541
 Thumbs Down       0.97      0.93      0.95       479
```

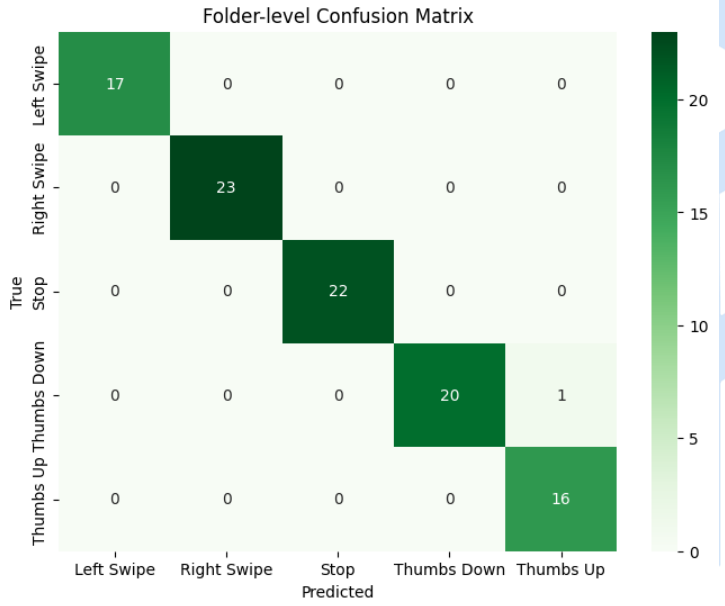
```
   Thumbs Up       0.95      0.94      0.95       363

    accuracy                           0.96      2061
   macro avg       0.96      0.96      0.96      2061
weighted avg       0.96      0.96      0.96      2061
```

گزارش طبقه‌بندی برای هر ژست به‌طور جداگانه دقت، یادآوری و امتیاز F1 را ارائه می‌دهد. در این بخش، تمامی ژست‌ها عملکرد خوبی داشتند، به‌ویژه ژست‌های "Left Swipe" و "Right Swipe" که دارای دقت بالایی بودند. دقت کلی گزارش نیز ٪۹۶ است.

## نتایج ارزیابی مدل با استفاده از رای‌گیری در سطح پوشه

برای مرحله بعدی ارزیابی، از رای‌گیری در سطح پوشه استفاده کردیم. در این روش، به ازای هر پوشه‌ای که شامل فریم‌های یک حرکت خاص بود، پیش‌بینی ژست انجام شده و بر اساس بیشترین تعداد پیش‌بینی‌ها، ژست نهایی برای آن پوشه انتخاب شد.




نتایج این روش نشان داد که مدل توانسته است ژست‌ها را با دقت ٪۹۸.۹۹ در سطح پوشه شناسایی کند. این به این معنی است که رای‌گیری در سطح پوشه به مدل کمک کرده تا پیش‌بینی‌های دقیق‌تری داشته باشد و اشتباهات احتمالی ناشی از خطاهای جزئی در پیش‌بینی‌های فریم‌های فردی را کاهش دهد.

```
Folder-level Accuracy: 0.9899

Folder-level Classification Report:
              precision    recall  f1-score   support

  Left Swipe       1.00      1.00      1.00        17
```

| | | | | |
|---|---|---|---|---|
| Right Swipe | 1.00 | 1.00 | 1.00 | 23 |
| Stop | 1.00 | 1.00 | 1.00 | 22 |
| Thumbs Down | 1.00 | 0.95 | 0.98 | 21 |
| Thumbs Up | 0.94 | 1.00 | 0.97 | 16 |
| | | | | |
| accuracy | | | 0.99 | 99 |
| macro avg | 0.99 | 0.99 | 0.99 | 99 |
| weighted avg | 0.99 | 0.99 | 0.99 | 99 |

در این بخش نیز عملکرد مدل بسیار عالی بود. تمامی ژست‌ها عملکرد خوبی داشتند، به‌ویژه برای ژست‌هایی مانند "Left Swipe" و "Right Swipe" که به‌طور کامل پیش‌بینی شدند. برای ژست "Thumbs Up" و "Thumbs Down" نیز مدل به خوبی عمل کرده بود.

### تحلیل نتایج

با مقایسه نتایج حاصل از دو روش مختلف ارزیابی (اسلایدینگ ویندو و رای‌گیری در سطح پوشه)، می‌توان مشاهده کرد که مدل در هر دو روش عملکرد بسیار خوبی داشته است. دقت بالای ٪۹۶ در ارزیابی با پنجره‌های ۵ فریمه و دقت بسیار بالا در ارزیابی با رای‌گیری در سطح پوشه (٪۹۸.۹۹) نشان‌دهنده قابلیت بالای مدل در شناسایی و دسته‌بندی ژست‌ها است. این نتایج نشان می‌دهد که مدل نه تنها می‌تواند ژست‌ها را در شرایط مختلف شناسایی کند، بلکه قادر است به خوبی از داده‌های زمان واقعی و متحرک نیز پیش‌بینی دقیقی ارائه دهد.

## بحث و نتیجه‌گیری

نتایج این پروژه نشان می‌دهند که استفاده از Mediapipe برای شناسایی دست‌ها و استفاده از یادگیری ماشین برای شناسایی ژست‌های دست می‌تواند عملکرد دقیقی را ارائه دهد. این سیستم می‌تواند در کاربردهای مختلف از جمله کنترل تلویزیون‌های هوشمند، سیستم‌های بازی، و حتی در محیط‌های صنعتی بدون نیاز به ورودی‌های فیزیکی استفاده شود. علاوه بر این، این سیستم به راحتی قابل گسترش به ژست‌های بیشتر است و می‌تواند به‌عنوان یک راه‌حل جذاب در تعاملات بدون تماس با دستگاه‌ها مورد استفاده قرار گیرد.

با توجه به نتایج خوب این پروژه، می‌توان آن را در دیگر زمینه‌های کاربردی نیز توسعه داد و به‌طور مؤثر در زندگی روزمره انسان‌ها استفاده کرد.

# مراجع

**[1]** M. Sagar, "Hand Gesture Detection System", *Kaggle*, 2021. Available: https://www.kaggle.com/datasets/marusagar/hand-gesture-detection-system/data. [Accessed: Sep. 2025].