بسمه تعالی
جمهوری اسلامی ایران

فرماندهی کل قوا
ستاد کل نیروهای مسلح
دانشگاه و پژوهشگاه عالی دفاع ملی و تحقیقات راهبردی

مرکز نخبگان واستعدادهای برتر نیروهای مسلح

« گزارش پروژه تحقیقاتی نخبگان »

عنوان طرح پژوهشي	مدل جفت شدگی جدید در رویدادهای دو یا سه الکترونی کوارک بالا در شتابدهنده LHC

گزارش دوم

کارگروه تخصصی: علوم پایه
نام و امضای نماینده سازمان کاربر طرح: مجتبی محمدی نجف آبادی
نام و امضای مجري طرح: میثم قاسمی بستان آباد
نام و امضای ناظر طرح: دکتر مجتبی محمدی نجف آبادی
طبقه‌بندي طرح: ذرات بنیادی
شماره و تاریخ نامه مصوبه طرح:

روکش گزارش

1) عنوان طرح که به تایید مرکز نخبگان رسیده است :

مدل جفت شدگی جدید در رویدادهای دو یا سه الکترونی کوارک بالا در شتابدهنده LHC.

2) هدف طرح که به تایید مرکز نخبگان رسیده است :

3) شرح خدمات و مراحل انجام و گام های تحقیق و جدول زمان بندی که به تایید مرکز نخبگان رسیده است:

ردیف	مراحل و گام‌هاي اجراي پروژه (توضیح مختصر در مورد هر مرحله و گام)	درصد مرحله (گام)	زمان اجرا (ماه)										
			1	2	3	4	5	6	7	8	9	10	11	12
1	مطالعه سیگنال و پس زمینه ها	٪۱۰	●											
2	تولید داده سیگنال و پس زمینه ها	٪۲۰			●									
3	محاسبه متغیرها با هوش مصنوعی	٪۳۰				●								
4	جداسازی سیگنال از پس زمینه ها	٪۲۰		●										
5	انجام تست های آماری	٪۲۰			●									
6	نگارش مقاله		●											

4) مختصری از گزارش قبلی، اصلاحات درخواست شده مرکز نخبگان و اصلاحات انجام شده

این گزارش اول میباشد.

5) خلاصه نتایج کسب شده در فاز جاری

سیگنال و پس زمینه ها بصورت دقیق مطالعه شده و تعداد مورد نیاز داده تولید شده است. در قدم بعد تمامی متغیرهای مربوط به ذرات نهایی محاسبه شده و تابع توزیع آنها ترسیم شده است. این متغیرها سپس بعنوان ورودی به مدل های هوش مصنوعی داده خواهند شد تا وزن های سیگنالی یا پس زمینه ای برای هر پدیده محاسبه گردد.

6) چنانچه از زمان بندی مصوب پروژه تاخیر دارد، علت تاخیر و چگونگی جبران تاخیر ذکر شود.

7) گزارش فاز جاری با فرمت و فصول ذکر شده در پیوست.

در صفحات بعدی ذکر شده است.

پیوست:

فصولی که در گزارش طرح پژوهشی در موضوعات فنی و مهندسی باید درج شوند

عنوان: شامل عنوان طرح، نام محقق، نام ناظر یا استاد راهنما، تاریخ و نام سازمان کارفرمای طرح
چکیده: شامل چکیده ای از اهمیت موضوع، کارهای دیگران، روش تحقیق، اهم نتایج بدست آمده و اهم تحیلی نتایج. چکیده حداکثر در دو صفحه است.
فهرست مطالب: فهرست مطالب در سه سطح ذکر شود. با رعایت روش نگارش اعلام شده، در تهیه متن از heading در سه سطح 1، 2 و 3 استفاده شود تا در نهایت با انتخاب Table of content نرم افزار خود، فهرست را تهیه کند.
مقدمه: شرح اهمیت موضوع انتخاب شده در حداکثر سه صفحه
مروری بر منابع: اعم از کتب، مقالات، ثبت اختراعات، سایت های اینترنتی معتبر، استانداردهای نظامی و غیرنظامی، دستورالعمل ها و ...
اهداف پروژه: بیان اهداف پروژه و علت انتخاب پروژه با توجه به کارهای انجام شده ذکر شده در بخش مروری بر منابع
روش تحقیق: شامل فلوچارت و توضیح فرایند تحقیق، مواد اولیه، روش دقیق انجام آزمایشات، توضیح نوع و مدل دستگاه های مورد استفاده و محلی که دستگاه مورد استفاده واقع شده است،
نتیجه آزمایشات: توضیح نتایج همراه با اشکال و نمودارها
تحلیل نتایج: با استفاده از نتایج بدست آمده و کمک از کارهای دیگران که در بخش مروری بر منابع آمده است، نتایج تحلیل شوند.
نتیجه گیری: میزان مطابقت نتیجه آزمایشات با اهداف پروژه بخوبی بیان شود.
منابع: فهرست منابعی که در فصول قبل استفاده شده است.
ضمائم و پیوست ها: چنانچه در انجام تحقیق از نرم افزار استفاده شده، نحوه استفاده از نرم افزار گفته شود.

چنانچه از روش تحقیق خاصی استفاده شده، روش در پیوست توضیح داده می شود. ضمائم و پیوست ها اجباری نیستند.

در تدوین گزارش های میانی و نهایی از فرمت زیر استفاده شود:

فونت‌ها	
عنوان طرح	B Lotus14
	Times New Roman 12
تیترهای اصلی متن	B Titr 14
زیرتیترها	B Titr 13
متن اصلی	B Lotus14
تیتر اشکال، جداول و نمودارهای داخل متن	B Titr 10
متن جداول	B Lotus12
ارجاعات فارسی پایین صفحه	B Lotus10
ارجاعات لاتین پایین صفحه	Times New Roman 10
فهرست منابع و مآخذ فارسی و عربی	B Zar14
فهرست منابع و مآخذ لاتین	Times New Roman 14

مرکز نخبگان واستعدادهای برتر نیروهای مسلح

عنوان: مدل جفت شدگی جدید در رویدادهای دو یا سه الکترونی کوارک بالا در شتابدهنده LHC

نام محقق: میثم قاسمی بستان آباد

نام ناظر یا استاد راهنما: دکتر مجتبی محمدی نجف آبادی

تاریخ: ۱۴۰۱/۳/۵

نام سازمان کارفرمای طرح: پژوهشگاه دانش های بنیادی

چکیده:

در این پروژه ما به دنبال یافتن اثرات تغییر طعم در کوارک بالا [۱] [۲] مانند تبدیل کوارک سر به کوارک بالا در برخوردهای پروتون - پروتون در شتابدهنده های هادرونی بزرگ در آزمایشگاه سرن میباشیم. این پدیده ها تنها در تصحیحات بالا در نظریه مدل استاندارد [۳] امکان پذیر میباشند. هرگونه (کشف احتمالی) سیگنال از تغییر طعم در بخش کوارک سر میتواند نشان دهنده طعم جدید لپتونی در فیزیک ماورأ مدل استاندارد باشد. این موضوع باعث شده تا تحقیقات گسترده ای در دو قالب تئوری و آزمایشگاهی در زمینه کشف تغییر طعم در آزمایشگاه های بزرگ دنیا از جمله ATLAS , CMS [۴] در LHC در سرن انجام شود. ذرات تشکیل دهنده سیگنال در این پروژه، سه لپتون با طعم یکسان و بار الکتریکی متفاوت، کوارک پایین و یک کوراک سبک میباشند. وجود دو یا سه لپتون با تکانه عرضی بالا و باردار و کوارک پایین امکان داشتن راندمان بالا (در بازسازی پدیده های سیگنالی) با استفاده از گیراندازی لپتون را فراهم مینماید. مهمترین پس زمینه های مدل استاندارد در این آنالیز عبارتند از: جفت کوارک سر (که از پدیده هایی مثل نابودی کوارک–ضدکوارک و همجوشی گلئون–گلئون میآیند. جفت کوارک سر سپس به سایر کانالها تبدیل میشوند: مانند دو لپتونی، تک لپتونی و تمام هادرونی)، رویداد های تک لپتون ناشی از واپاشی کوارک سر، و تک لپتون کوارک سر به همراه بوزون Z یا W. شبیه سازهای مونت کارلو در این پروژه برای تولید داده های سیگنال و پس زمینه استفاده شده اند. پدیده های آبشار پارتونی، رویدادهای زمینه ای و برهم کنش ذرات نهایی با ابعاد آشکارساز تماما با شبیه سازهای تخصصی شبیه سازی شده اند. برای جدا کردن سیگنال از پس زمینه های نظریه مدل استاندارد ، میتوان از انتخابات سه لپتونی به همراه کوارک پایین و یا برش پنجره ای بر روی جرم کوارک سر استفاده کرد. روشهای متعددی برای کاهش دادن بیشتر پس زمینه های احتمالی و افزایش راندمان سیگنال وجود دارد که میتوان به بکارگیری وزن های هوش مصنوعی حاصله از الگوریتم درختی یا شبکه عصبی اشاره کرد. از دیگر موارد برای بهبود آنالیز، تعریف ناحیه های حساس به سیگنال برای سناریوهای اسکالر، برداری و تنسوری میباشد. در قدم نهایی نیاز میباشد تا مقادیر پی برای سیگنال و پس زمینه ها اندازه گیری و سپس با استفاده از روش های تست آماری مقادیر ممنوعه با احتمال ۹۵٪ برای مقیاس جرمی فیزیک جدید گزارش شود. هرگونه کشف احتمالی در این تحقیق به درک عمیقتر ما از تغییر طعم در کوارک بالا منجر میشود و نتایج این پروژه میتواند در دیگر آنالیزهای ماورأی مدل استاندارد مورد استفاده قرار بگیرد.

۱. مقدمه:

درآزمایش‌های فیزیک انرژی بالا (HEP) مقادیر زیادی داده تولید می شوند و این موضوع توصیف برداشت های آماری معنادار و کشف ذرات یا پدیده‌های جدید را به چالش می‌کشد. تکنیک‌های یادگیری ماشینی (ML) به عنوان ابزار قدرتمندی برای مقابله با این سیل داده‌ها و استخراج مدل های آماری ظهور کرده‌اند. در ذیل بخشی از کاربردهای ML را در جنبه‌های مختلف تحقیقات HEP ، از جمله تجزیه و تحلیل داده‌ها، بازسازی رویداد، تشخیص ناهنجاری و شناسایی ذرات بررسی می‌کنیم.

- معرفی

آزمایش‌های فیزیک انرژی بالا، مانند آزمایش‌هایی که در برخورد دهنده بزرگ هادرون (LHC) انجام می شود، داده‌های بسیار زیادی از برخورد ذرات تولید می‌کند. روش های سنتی تجزیه و تحلیل داده ها با چالش های قابل توجهی در مقابله با پیچیدگی و حجم این داده ها روبرو هستند. تکنیک‌های یادگیری ماشینی یک رویکرد جایگزین ارائه می‌کنند و محققان را قادر می‌سازد تا بینش‌های ارزشمندی را از مجموعه داده‌های گسترده استخراج و مدل های مورد نظر را در زمینه های مختلف استفاده کنند.

- تحلیل داده ها (Data Analysis)

یکی از کاربردهای اولیه ML در HEP تجزیه و تحلیل داده ها است. الگوریتم‌های ML مانند شبکه‌های عصبی عمیق و تصمیمات درختی را می‌توان برای تحلیل داده‌های برخورد ذرات و شناسایی الگوها یا سیگنال‌های خاص آموزش داد. با استفاده از تکنیک‌های یادگیری نظارت شده، محققان می‌توانند مدل‌های ML را برای طبقه بندی ذرات، تمایز سیگنال از پس‌زمینه و تخمین خواص ذرات تازه کشف‌شده آموزش دهند. الگوریتم‌های ML عملکرد بسیار خوبی در تمایز بین اثرات ذرات مختلف و افزایش دقت کلی تحلیل داده‌ها نشان داده‌اند.

- بازسازی رویداد (Event Reconstrution)

بازسازی رویداد یک مرحله حیاتی در آزمایش‌های HEP است که در آن داده‌های خام آشکارساز برای بازسازی مسیرها و ویژگی‌های ذرات تولید شده در یک برخورد پردازش می‌شوند. تکنیک های ML می توانند کارایی و دقت الگوریتم های بازسازی رویداد را به طور قابل توجهی بهبود بخشند. با آموزش مدل‌ها بر روی داده‌های شبیه سازی شده یا استفاده از تکنیک‌های یادگیری بدون نظارت، الگوریتم‌های ML می‌توانند الگوهای پیچیده‌ای را در پاسخ آشکارساز بیاموزند، خطاهای بازسازی را کاهش داده و کیفیت کلی بازسازی را افزایش دهند.

- تشخیص ناهنجاری (Anomaly Detection)

تشخیص ناهنجاری نقشی مهمی در آزمایش‌های HEP ایفا می‌کند، زیرا به شناسایی رویدادهای نادر که می‌توانند حضور پدیده‌های جدید فیزیک را نشان دهند، کمک می‌کند. الگوریتم‌های ML ، مانند رمزگذارهای خودکار یا شبکه های مولد [۵] (GANs) ، می‌توانند برای یادگیری رفتار عادی آزمایش و تشخیص انحرافات

از آن استفاده شوند. این مدل‌ها می‌توانند سیگنال‌های ظریفی را شناسایی کنند که ممکن است با روش‌های تحلیل سنتی نادیده گرفته شوند و امکان کشف ذرات یا برهمکنش های جدید را فراهم کنند.

- شناسایی ذرات (Particle Identification)

تمایز بین انواع مختلف ذرات یک کار اساسی در HEP است. تکنیک‌های ML در آنالیزهای شناسایی ذرات بسیار مؤثر بوده‌اند. شبکه های عصبی عمیق [۶] (DNN) و یا پیچیده [۷] (convolution) را می توان بر روی داده های شبیه سازی شده برای شناسایی انواع ذرات خاص بر اساس انرژی ذرات و پاسخ آشکارساز آموزش داد. علاوه بر این، روش‌های تقویت گرادیان (Gradiant boosting) و ماشین‌های بردار پشتیبان (SVM)را می‌توان برای کارهای طبقه‌بندی ذرات، دستیابی به دقت و استحکام بالا مورد استفاده قرار داد.

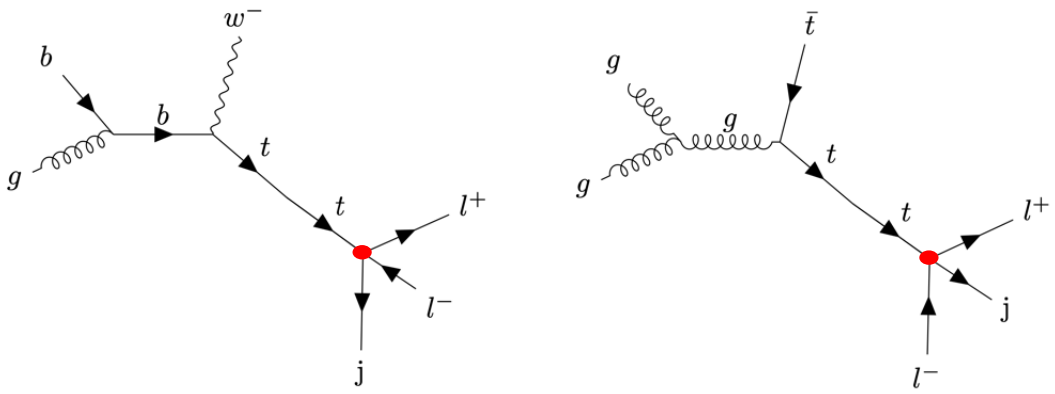
- نتیجه

تکنیک های یادگیری ماشین به ابزارهای ارزشمندی در زمینه فیزیک انرژی بالا تبدیل شده اند. آنها قابلیت های قدرتمندی برای تجزیه و تحلیل داده ها، بازسازی رویداد، تشخیص ناهنجاری و شناسایی ذرات ارائه می دهند. با استفاده از این تکنیک‌ها، محققان می‌توانند دقت، کارایی و پتانسیل کشف آزمایش‌های HEP را بهبود بخشند. پیشرفت های مداوم در الگوریتم های ML توسط مهندسین، بدون شک آینده تحقیقات فیزیک انرژی بالا را شکل خواهد داد. در این گزارش نتایج مختلف مدل های هوش مصنوعی در شناسایی ذرات سیگنال و پس زمینه با استفاده از داده های شبیه سازی شده شرح داده میشود.

۲. اهداف گزارش :

يكي از مدل‌هاي موفق در فيزيك ذرات با انرژي بالا، مدل استاندارد ذرات است. شتاب‌دهنده بزرگ پروتون – پروتون LHC واقع در CERN كه در محدوده انرژي ترا الكترون ولت TeV كار مي‌كند آزمون‌هاي تجربي بسياري را بر روي مدل استاندارد ذرات انجام داده است. ساختار كلي برخورد دهنده بزرگ هادروني (LHC) بصـــورت برخورد دهنده‌اي دايروي اســـت كه دو باريكه پروتون را با انرژي TeV ۷+۷ به هم مي‌كوبد. تونل LHC به طول ۲۷ كيلومتر است كه در ۴ نقطه از آن دسته‌هاي پروتون با انرژي مركز جرم ۱۴TeV به هم برخورد مي‌كنند و امكان توليد ذرات سـنگين‌تر را فراهم مي‌آورند. در این پروژه ما به دنبال تغییر طعم در کوارک بالا مانند تبدیل کوارک سـر به کوارک بالا میباشـیم. این پدیده ها تنها در تصـحیحات بالا در نظریه اسـتاندارد مدل امکان پذیر میباشـند. هرگونه (کشـف احتمالی) سـیگنال از تغییر طعم در بخش کوارک سر میتواند نشان دهنده طعم جدید لپتونی در فیزیک ماورأ استاندارد مدل باشد.

بعنوان یادآوری ذرات تشـکیل دهنده سـیگنال در این پروژه سـه لپتون با بار الکتریکی متفاوت، یک کوارک پایین و یک کوارک سبک میباشند. شکل ۱ نمودار فاینمن سیگنال مورد مطالعه را نشان میدهد که l نشانگر الکترون در ذرات نهایی است.



شکل ۱. نمودار فاینمن سیگنال با تغییر طعم کوارک سر. شکل سمت راست سیگنال tt و شکل سمت چپ سیگنال tW میباشند. تنها الکترون ها بعنوان لپتون در این آنالیز مورد بررسی قرار گرفته اند. راس قرمز رنگ نشان دهنده راس مدل غیر استاندارد برای واپاشی کوارک سر میباشد.

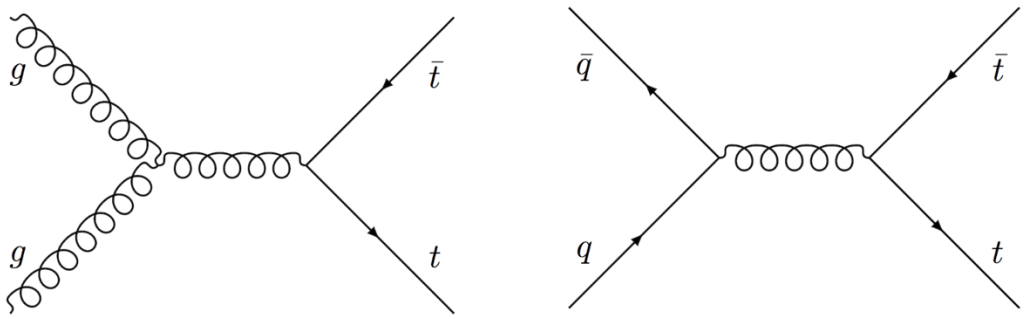
وجود دو یا سه لپتون با تکانه عرض بالا (P_T) و باردار و کوارک پایین امکان داشتن راندمان بالا با استفاده از گیراندازی لپتون (di-lepton trigger) فراهم مینماید. مهمترین پس زمینه های مدل استاندارد در این آنالیز عبارتند از:

- جفت کوارک سر که از پدیده هایی مثل نابودی کوارک – ضدکوارک و همجوشی گلئون – گلئون تولید میشوند. جفت کوارک سر سپس به سایر کانالها تبدیل میشود: مانند دو لپتونی، تک لپتونی و تمام هادرونی. یکی از لپتون ها در این پس زمینه بصورت جعلی میباشد مانند لپتون جعلی از تابش

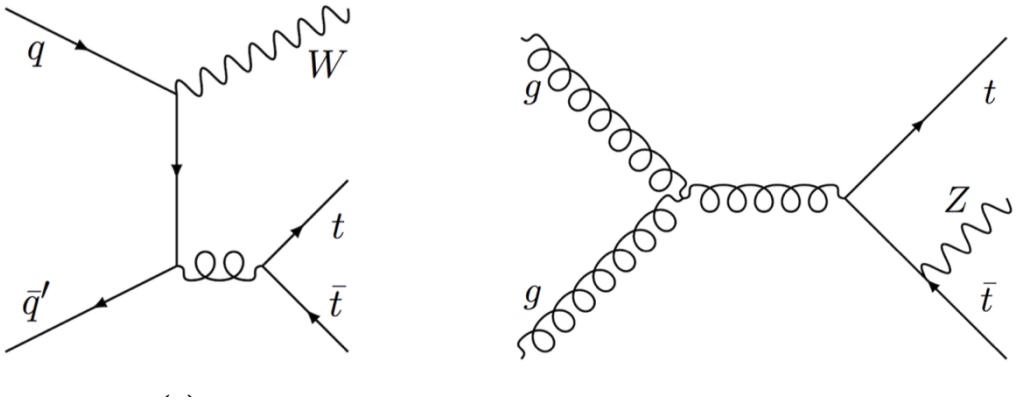
و یا جتی که بصورت الکترون بازسازی شده است. شکل ۲ نمودار فاینمن تولید جفت کوارک سر را نشان میدهد.

- جفت کوارک سر به همراه بوزون W یا Z. در این مدل پس زمینه، بوزون ها به صورت لپتونی واپاشی میکنند تا تعداد ۳ الکترون در فضای نهایی داشته باشیم. شکل ۳ نمودار فاینمن تولید جفت کوارک سربه همراه بوزون W یا Z را نشان میدهد.
- تک کوارک سر به همراه بوزون Z که بوزون به صورت لپتونی واپاشی میکند.

از دیگر پس زمینه ها میتوان به چهار کوارک سر و دو بوزونی مانند WZ که بصورت لپتونی واپاشی میکنند اشاره کرد. هدف اصلی این پروژه تولید سیگنال و پس زمینه ها و جداسازی انها به شکل مطلوب با استفاده از روش های هوش مصنوعی برای انجام تست های آماری میباشد. برای رسیدن به این هدف در ابتدا داده های سیگنالی و پس زمینه ای بعنوان ورودی مدلها مورد استفاده قرار گرفته تا فرآیند یادگیری به شکل مطلوب انجام گیرد. سپس وزن های مدل ها بعنوان جداسازنده (Discriminator) نقش مهمی در تعریف نواحی حساس سیگنالی خواهند داشت.



شکل ۲. نمودار فاینمن پس زمینه جفت کوارک سر. شکل سمت راست نابودی کوارک – ضدکوارک و شکل سمت چپ همجوشی گلئون – گلئون را نشان میدهند.



شکل ۳. نمودار فاینمن پس زمینه جفت کوارک سر به همراه بوزون W یا Z.

۳. کاربرد هوش مصنوعی در این آنالیز:

تکنیک‌های یادگیری ماشین ثابت کرده‌اند که در کار تشخیص سیگنال و پس‌زمینه در آزمایش‌های فیزیک انرژی بالا بسیار مؤثر هستند. تمایز بین رویدادهای سیگنال، که حاوی اطلاعات ارزشمندی در مورد پدیده‌های بالقوه فیزیک جدید هستند، و رویدادهای پس‌زمینه، که از فرآیندهای شناخته شده ناشی می‌شوند، برای استخراج بینش معنادار از داده‌ها بسیار مهم است. الگوریتم‌های ML، مانند شبکه‌های عصبی عمیق (DNN)، قابلیت‌های قابل‌توجهی در وظایف تشخیص سیگنال و پس‌زمینه نشان داده‌اند. با آموزش بر روی داده های برچسب گذاری شده، DNN ها می توانند الگوها و همبستگی های پیچیده (correlations) در داده ها را بیاموزند و آنها را قادر می سازد بین رویدادهای سیگنال و پس زمینه به طور دقیق تمایز قائل شوند. مزیت DNN ها در توانایی آن ها برای استخراج خودکار ویژگی های مربوطه از داده های ورودی نهفته است. همچنین این مدل می‌تواند همزمان، استفاده از متغیرهای زیادی را که سیگنال و پس‌زمینه را جدا می‌کنند، بهینه کند. یک تکنیک یادگیری ماشینی نه تنها استفاده از متغیرهای زیادی را به طور همزمان بهینه می‌کند، بلکه می‌تواند همبستگی‌هایی را در ابعاد مختلف پیدا کند، که نسبت به متغیرهای جداگانه طبقه‌بندی سیگنال/پس‌زمینه بهتری را ارائه می‌دهد.

علاوه برDNN ها، سایر الگوریتم‌های ML مانند جنگل‌های تصادفی [۸] و روش‌های تقویت گرادیان با موفقیت در وظایف تشخیص سیگنال و پس‌زمینه استفاده شده‌اند. این الگوریتم‌ها از تکنیک‌های یادگیری گروهی استفاده می‌کنند و می‌توانند فضاهای ویژگی با ابعاد بالا را مدیریت کنند. با ترکیب چند طبقه‌بندی‌کننده ضعیف، آنها می‌توانند به طور موثر مرزهای تصمیم پیچیده بین رویدادهای سیگنال و پس‌زمینه را مدل‌سازی کنند. این الگوریتم‌ها در گرفتن روابط غیرخطی عالی هستند و می‌توانند متغیرهای پیوسته و طبقه‌ای را مدیریت کنند و آنها را برای تمایز بین امضاهای ذرات مختلف مناسب می‌سازد.

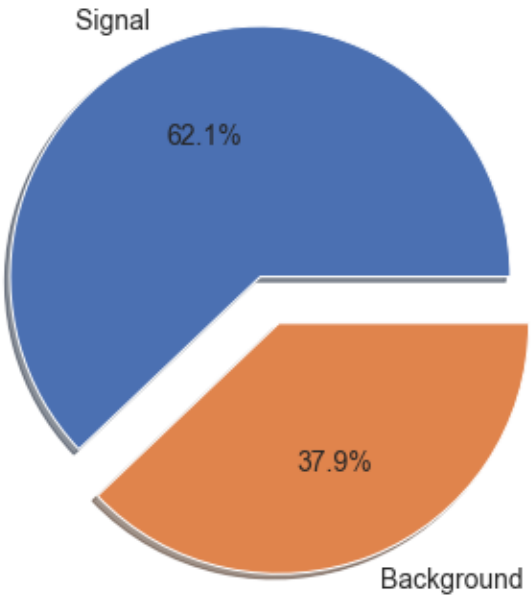
برای اطمینان از قابلیت تعمیم و استحکام مدل‌هایML ، اعتبارسنجی و آزمایش گسترده با استفاده از مجموعه داده‌های مستقل انجام می‌شود. تکنیک‌های اعتبارسنجی متقابل، مانند اعتبارسنجی متقاطع k-fold ، معمولاً برای ارزیابی عملکرد مدل‌ها استفاده می‌شوند. علاوه بر این، محققان از تکنیک‌هایی مانند منحنی‌های مشخصه عملکرد گیرنده (ROC) و ناحیه زیر منحنی (AUC) برای ارزیابی و مقایسه عملکرد تمایز مدل‌های مختلف ML استفاده می‌کنند. در نتیجه، تکنیک های یادگیری ماشین، انقلابی در جداسازی سیگنال و پس زمینه در فیزیک انرژی بالا ایجاد کرده است. از طریق استفاده از الگوریتم‌های پیشرفته ML، مانند شبکه‌های عصبی عمیق و روش‌های دیگر، محققان می‌توانند به طور موثر رویدادهای سیگنال را از رویدادهای پس‌زمینه جدا کنند و امکان کشف پدیده‌های جدید فیزیک را فراهم کنند.

در این آنالیز به منظور یادگیری مدل های هوش مصنوعی، داده های سیگنالی و پس زمینه ای به نسبت ۶۰ به ۴۰ تقسیم شده و تنها پس زمینه های مهم (leading backgrounds) استفاده شده اند. این پس زمینه ها عبارتند از: $\bar{t}t, WZ, ZZ$. شکل ۴ نسبت این داده ها را نشان میدهد. در یادگیری ماشینی، ویژگی‌های ورودی (input features) نقش مهمی در آموزش مدل‌ها و پیش‌بینی‌ها یا طبقه‌بندی‌های دقیق دارند. انتخاب و درک

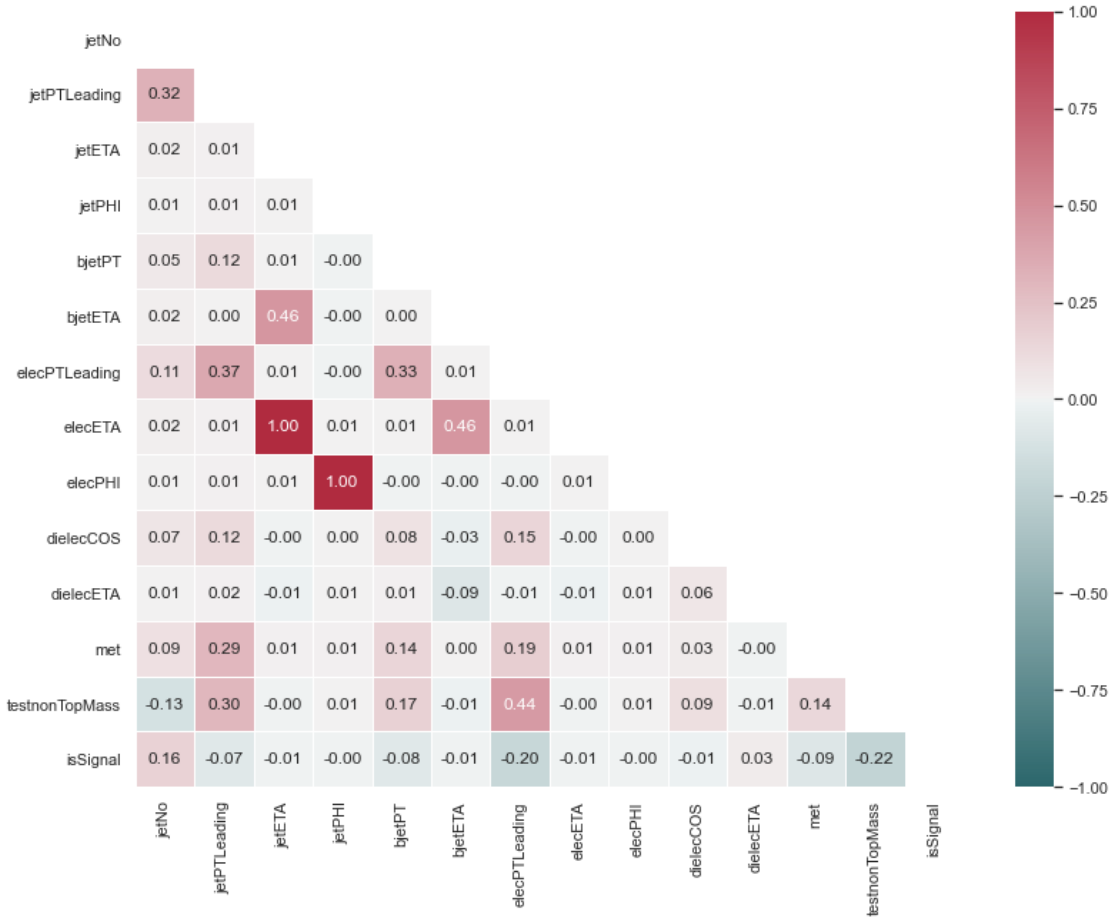
ویژگی های ورودی برای گرفتن اطلاعات مربوطه از داده ها و ساخت مدل های موثر ضروری است. علاوه بر این، درک همبستگی (correlation) بین ویژگی های ورودی برای جلوگیری از مسائلی مانند چند خطی بودن، که در آن ویژگی ها به شدت با یکدیگر همبستگی دارند، مهم است. همبستگی بالا می تواند منجر به بی ثباتی و مشکلات در تفسیر ضرایب مدل یا رتبه بندی اهمیت ویژگی شود. در چنین مواردی، مدل ممکن است بر برخی ویژگی‌ها بیش از حد تأکید کند، در حالی که برخی دیگر را که به همان اندازه آموزنده هستند، نادیده بگیرد. برای پرداختن به این موضوع، تحلیل همبستگی ویژگی‌ها را می‌توان انجام داد، مانند محاسبه ضرایب همبستگی یا تجسم روابط ویژگی با استفاده از تکنیک‌هایی مانند نمودارهای پراکندگی (scattering-plots) یا نقشه‌های حرارتی (heatmap).

شکل ۵ نمودار نقشه های حرارتی ویژگی های ورودی در این آنالیز را نشان میدهد (برای سیگنال کوارک c). سطر آخر isSignal بیانگر سیگنال یا پس زمینه بودن داده مورد نظر میباشد. مهمترین ویژگی ها با بالاترین ضریب همبستگی نسبت به سطر isSignal، جرم کوارک سر از راس غیر مدل استاندارد Topmass با ضریب 0.22- و تعداد جت jetNo با ضریب 0.16 میباشند. ضریب همبستگی منفی برای جرم کوارک سر بیانگر این است که با بالا رفتن مقدار جرم کوارک سر، احتمال پس زمینه بودن داده مورد نظر بالاتر میباشد. همچنین ضریب همبستگی مثبت برای تعداد جت، نشان دهنده بالاتر بودن احتمال داده سیگنالی در داشتن تعداد جت بیشتر میباشد. نمودار توزیع تمام ویژگی های ورودی در شکل ۶ نمایش داده شده است که خود تاییدیه بر بالاتر بودن مقدار میانگین برای داده های سیگنالی در متغیر جرم کوارک سر و پایین تر بودن مقدار میانگین برای تعداد جت میباشد (این موضوع در گزارش قبلی مورد بحث و بررسی قرار گرفته شده است).

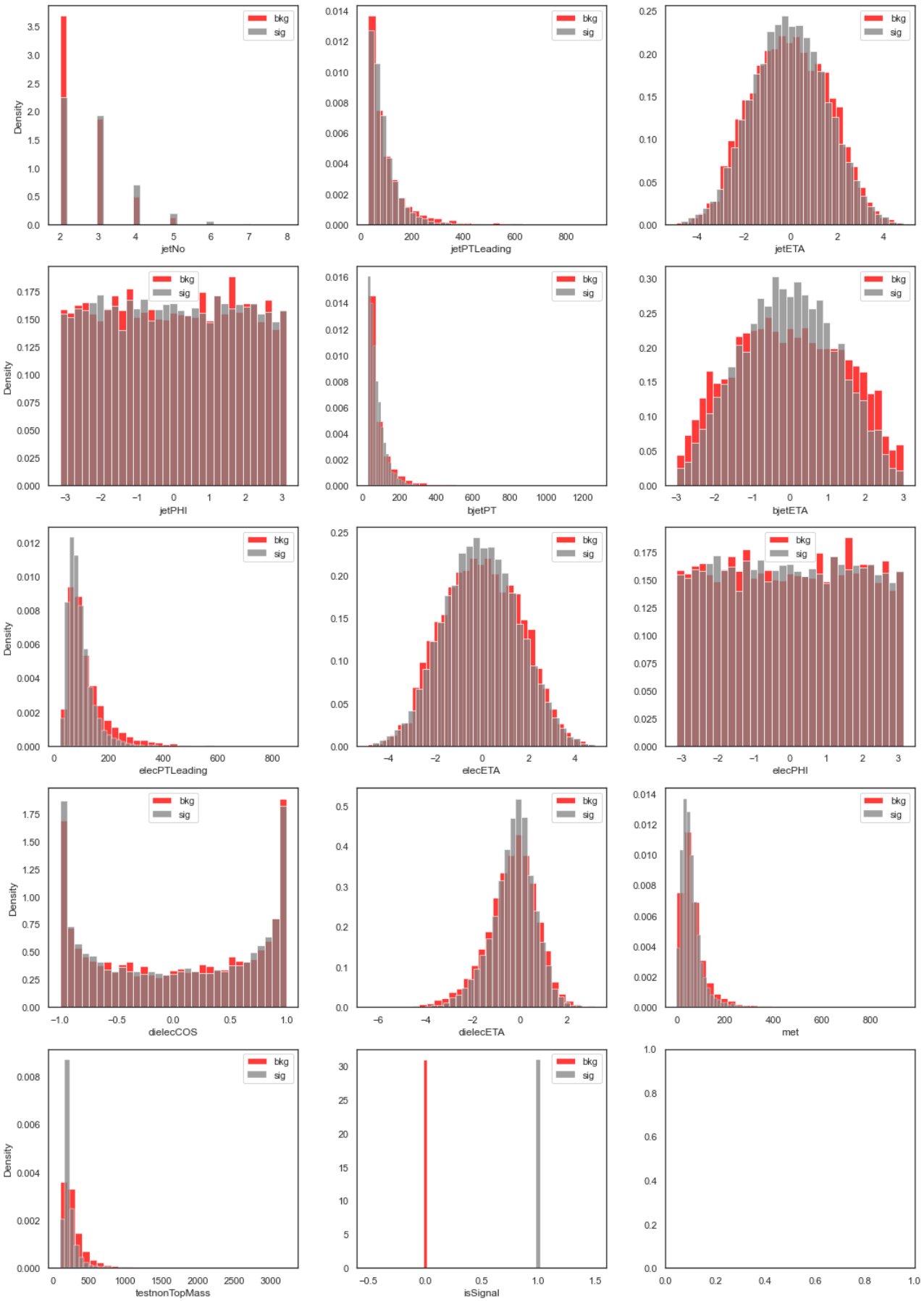
در انتها باید گفت یک تکنیک پیش پردازش متداول در یادگیری ماشینی که به استانداردسازی یا عادی سازی ویژگی های یک مجموعه داده کمک می کند StandardScaler نامیده میشود. فرآیند استانداردسازی شامل تفریق مقدار میانگین هر ویژگی و سپس تقسیم آن بر انحراف استاندارد است. این موضوع تضمین می کند که همه ویژگی ها دارای مقیاس های مشابه هستند و در مرکز صفر قرار دارند. استاندارد کردن ویژگی‌ها به ویژه هنگام کار با الگوریتم‌هایی مفید است که توزیع گاوسی (عادی) را فرض می‌کنند یا بر فاصله اقلیدسی بین نقاط داده تکیه می‌کنند، مانند ماشین‌های بردار پشتیبان (SVM) ، نزدیک‌ترین همسایگان (KNN)، یا مدل‌های رگرسیون خطی. استانداردسازی تمام ویژگی‌های ورودی در این آنالیز با استفاده از کتابخانه sikit-learn انجام شده است.



شکل ۴. نسبت داده های سیگنالی (۶۰٪) به پس زمینه ها (۴۰٪) به منظور ورودی برای یادگیری مدل های هوش مصنوعی. تمامی پس زمینه های مهم $(\bar{t}t, WZ, ZZ)$ با برچسب Background نمایش داده شده اند.



شکل ۵. نمودار نقشه های حرارتی ویژگی های ورودی برای سیگنال کوارک c. هر سلول نمایانگر ضریب همبستگی بین دو متغیر میباشد.



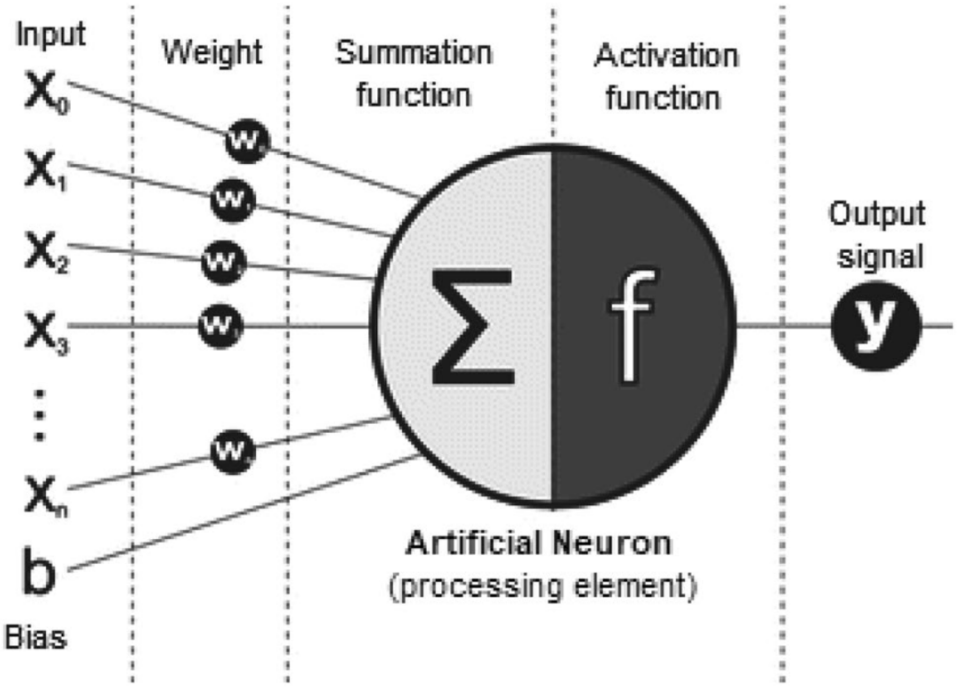
شکل ۶. نمودار توزیع ویژگی های ورودی برای سیگنال کوارک c (رنگ خاکستری) و پس زمینه ها (رنگ قرمز). متفارت بودن توزیع داده های سیگنالی و پس زمینه ای در متغیرهای تعداد جت و جرم کوارک سر، نشانگر قدرت بالای این متغیرها در جداسازی بین داده ها میباشد.

۳. نتایج مدل های هوش مصنوعی:

- شبکه عصبی عمیق (DNN)

شبکه های عصبی به عنوان یک رویکرد قدرتمند و همه کاره برای حل مسائل پیچیده محاسباتی ظهور کرده اند. این شبکه ها با تقلید از ساختار به هم پیوسته نورون های مغز انسان، قابلیت های قابل توجهی در یادگیری از داده ها، تشخیص الگوها و پیش بینی نشان داده اند. یک شبکه عصبی از لایه های به هم پیوسته نورون های مصنوعی یا «گره ها» تشکیل شده است. هر نورون ورودی ها را دریافت می کند، تبدیل های ریاضی را اعمال می کند و خروجی تولید می کند که به لایه بعدی ارسال می شود (شکل ۷). لایه ها معمولاً به سه نوع سازماندهی می شوند: لایه ورودی، لایه های پنهان و لایه خروجی. اتصالات بین نورون ها با وزن های قابل تنظیم مشخص می شود که قدرت و تأثیر هر اتصال را تعیین می کند.

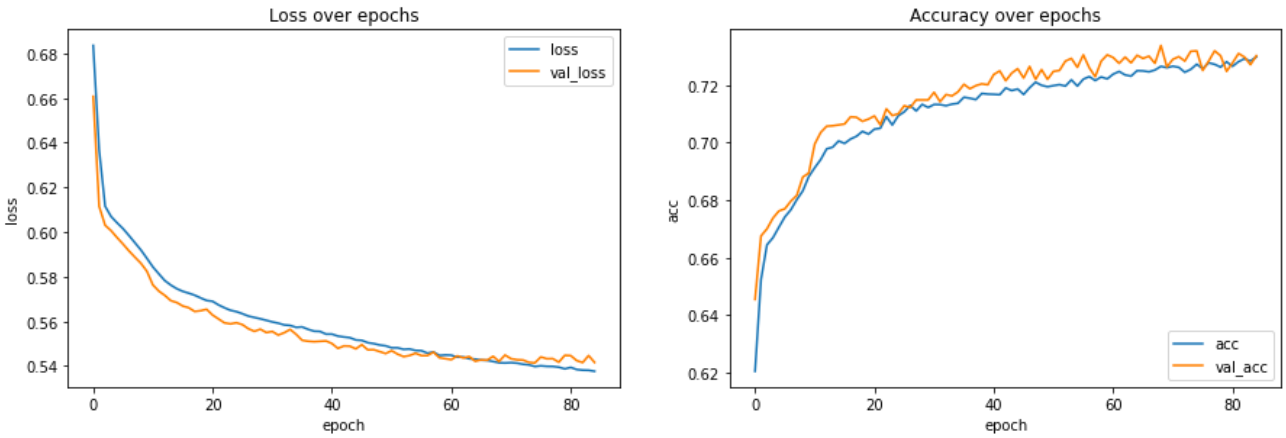
شبکه های عصبی از طریق فرآیندی به نام آموزش از داده ها یاد می گیرند. متداول‌ترین تکنیک، که به عنوان یادگیری نظارت شده شناخته می‌شود، شامل ارائه شبکه با نمونه‌های برچسب‌گذاری شده و تنظیم وزن‌ها برای به حداقل رساندن اختلاف بین خروجی‌های پیش‌بینی‌شده و برچسب‌های واقعی است. سایر تکنیک های یادگیری، از جمله یادگیری بدون نظارت، و یادگیری تقویتی نیز مورد استفاده قرار می گیرند. شبکه‌های عصبی عمیق، که با لایه‌های متعدد و معماری پیچیده‌شان مشخص می‌شوند، انقلابی در حوزه هوش مصنوعی ایجاد کرده‌اند. از جمله کاربردهای مختلف یادگیری عمیق را میتوان به بینایی رایانه، پردازش زبان طبیعی، تشخیص گفتار و سیستم‌های مستقل اشاره کرد. عملکرد استثنایی شبکه‌های عصبی عمیق در این حوزه‌ها، فرصت‌های جدیدی را ایجاد کرده و باعث پیشرفت‌هایی در زمینه‌هایی مانند مراقبت‌های بهداشتی، مالی، روباتیک و غیره شده است.



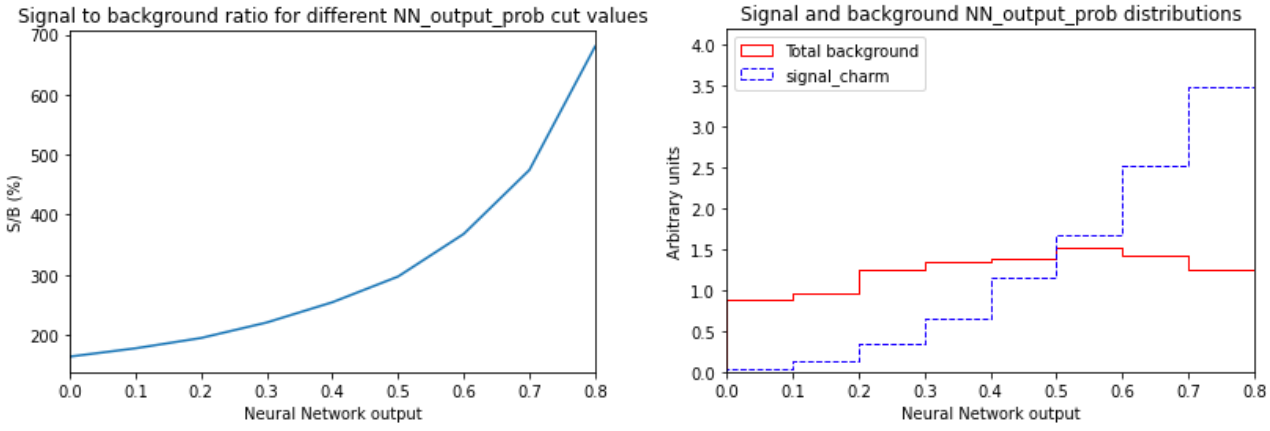
شکل ۷. صورت کلی ساختار شبکه عصبی به همراه لایه ورودی، لایه های میانی، وزن های اتصالات و لایه خروجی. در روند یادگیری با تنظیم وزن‌ها، اختلاف بین خروجی‌های پیش‌بینی‌شده و برچسب‌های واقعی به حداقل رسانده میشوند.

به منظور جداسازی سیگنال از پس زمینه ها در این آنالیز، دو مدل مختلف از شبکه عصبی مورد استفاده قرار گرفته شده است. مدل اول مدل ساده شبکه عصبی به همراه دو لایه میانی با ۲۰ گره و بدون بهینه سازی ابرپارامترهای مدل میباشد. به منظور جلوگیری از یادگیری بیش از حد overfit فرآیند یادگیری در نقطه ای که دقت مدل در داده های تستی test dataset کاهش یابد، مترقف میشود (روش early stopping). عملکرد مدل شبکه عصبی ساده بر روی داده های آموزشی و تستی در شکل ۸ نمایش داده شده است. محور افقی epoch نشان دهنده تعداد دفعاتی است که فرآیند یادگیری بر روی داده های آموزشی انجام میگیرد. محور عمودی در شکل سمت چپ loss تفاوت مقدار واقعی و مقدار پیش بینی شده توسط مدل و Accuracy نسبت دفعاتی که مدل درست پیشگویی کرده به کل تعداد داده ورودی است.

شکل ۹ (سمت راست) میزان اثر بخشی مدل شبکه عصبی ساده در جداسازی سیگنال و پس زمینه را نشان میدهد. با بالاتر رفتن قدرت مدل در جداسازی، نسبت سیگنال به پس زمینه (S/B) افزایش یافته و امکان تعریف ناحیه سیگنالی را به وجود می آورد (سمت چپ).



شکل ۸. مقادیر تفاوت loss (سمت چپ) و دقت accuracy (سمت راست) برای داده های آموزشی و تستی در هر دوره آموزش epoch. با تکرار روند آموزش مقدار پیش بینی شده توسط مدل به مقدار واقعی نزدیک شده و دقت افزایش می یابد.

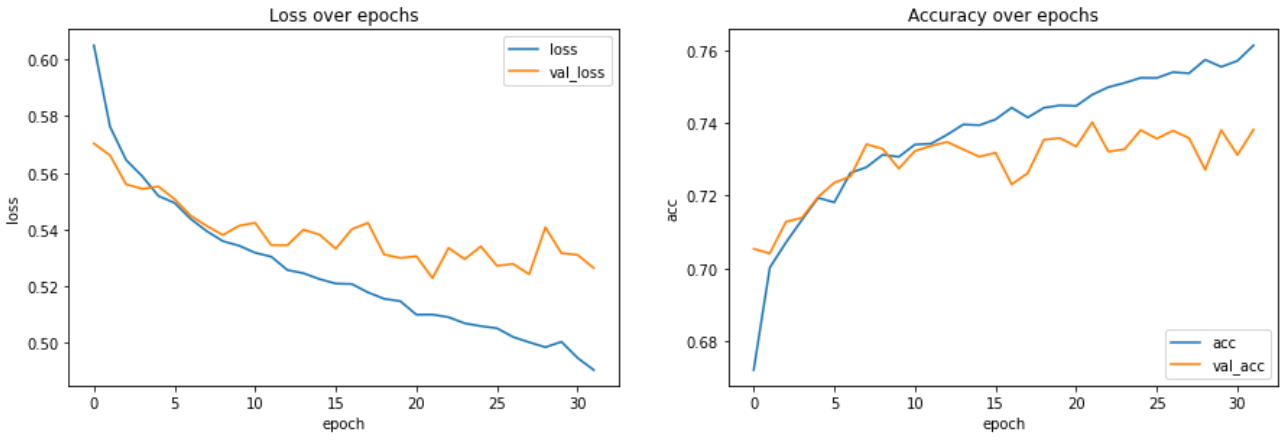


شکل ۹. نسبت تعداد داده های سیگنال به پس زمینه (S/B) و نمودار توزیع خروجی شبکه عصبی (احتمال سیگنال یا پس زمینه بودن) آنها.

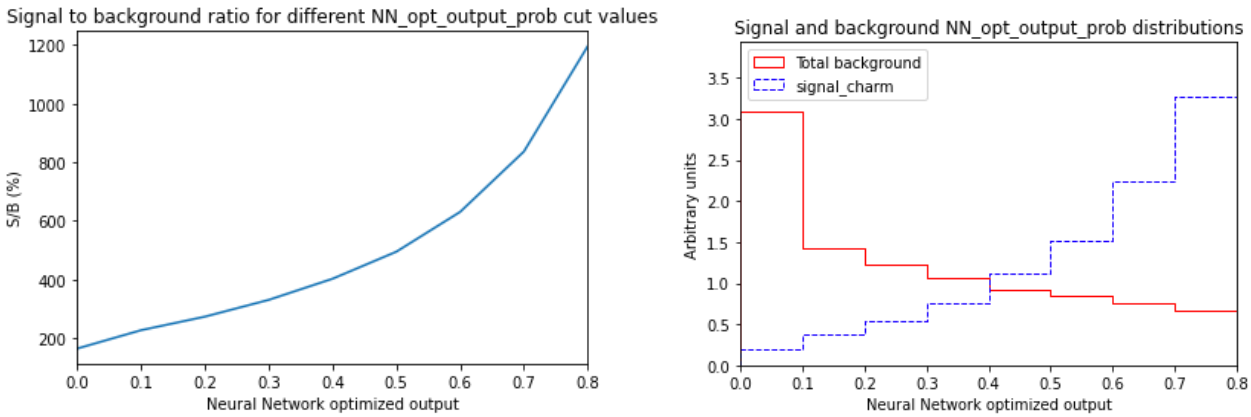
مدل دوم شامل بهینه سازی هایپرپارامترهای شبکه عصبی با استفاده از Keras Tuner است که به ما این امکان را می دهد به طور خودکار ترکیب بهینه ابرپارامترها را برای مدل شبکه عصبی خود جستجو کنیم. هایپرپارامتر در شبکه عصبی تنظیمات پیکربندی هستند که ساختار و رفتار شبکه عصبی را تعریف می‌کنند، مانند تعداد لایه‌ها، تعداد نورون‌ها در هر لایه، نرخ یادگیری و توابع فعال‌سازی. فرآیند بهینه سازی هایپرپارامتر با استفاده از Keras Tuner معمولاً شامل تعریف فضای جستجو (مقادیر احتمالی برای تعداد نورونها و یا نرخ یادگیری) و انتخاب الگوریتم جستجو RandomSearch میباشد. این الگوریتم یک جستجوی تصادفی را در فضای تعریف شده انجام می دهد. پس از تکمیل فرآیند جستجو، می توان نتایج را برای شناسایی بهترین پیکربندی هایپرپارامتر و معیارهای عملکرد مرتبط با آن بررسی کرد. تنظیم کننده (tunner) اطلاعاتی مانند بهترین مجموعه هایپرپارامترها، بهترین معماری مدل و نمرات عملکرد مربوطه را ارائه می دهد.

عملکرد مدل شبکه عصبی بهینه شده بر روی داده های آموزشی و تستی در شکل ۱۰ نمایش داده شده است. همان طور که از شکل پیداست، مقدار دقت شبکه عصبی بهینه شده بالاتر از شبکه عصبی ساده میباشد. شکل ۱۱ (سمت راست) میزان اثر بخشی مدل شبکه عصبی بهینه شده در جداسازی سیگنال و پس زمینه را نشان میدهد. با بالاتر رفتن قدرت مدل در جداسازی، نسبت سیگنال به پس زمینه (S/B) افزایش یافته و امکان تعریف ناحیه سیگنالی را به وجود می آورد (سمت چپ). میزان جداشدگی سیگنال از پس زمینه و نسبت آنها پس از بهینه سازی افزایش پیدا کرده است.

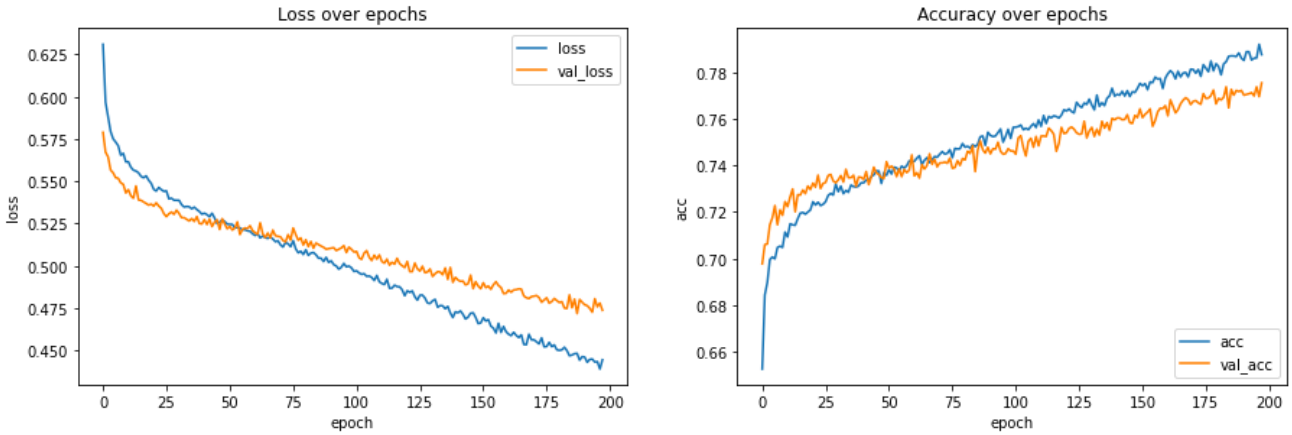
در حالتی دیگر میتوان برای جلوگیری از یادگیری بیش از حد مدل شبکه عصبی بهینه شده از لایه های حذفی (dropout) با نرخ ۱۰٪ استفاده کرد. در یک لایه حذفی، بصورت تصادفی کسری از واحدهای ورودی (نورون‌ها) در طول هر تکرار آموزشی، حذف می‌شوند یا به طور موقت نادیده گرفته می‌شوند. این به این معنی است که مشارکت آنها در گذررو به جلو forward propagation و رو به عقب backward propagation شبکه به طور موقت حذف می شود. اشکل ۱۲ و ۱۳عملکرد مدل شبکه عصبی بهینه شده با لایه های حذفی بر روی داده های آموزشی و تستی را نشان میدهد.



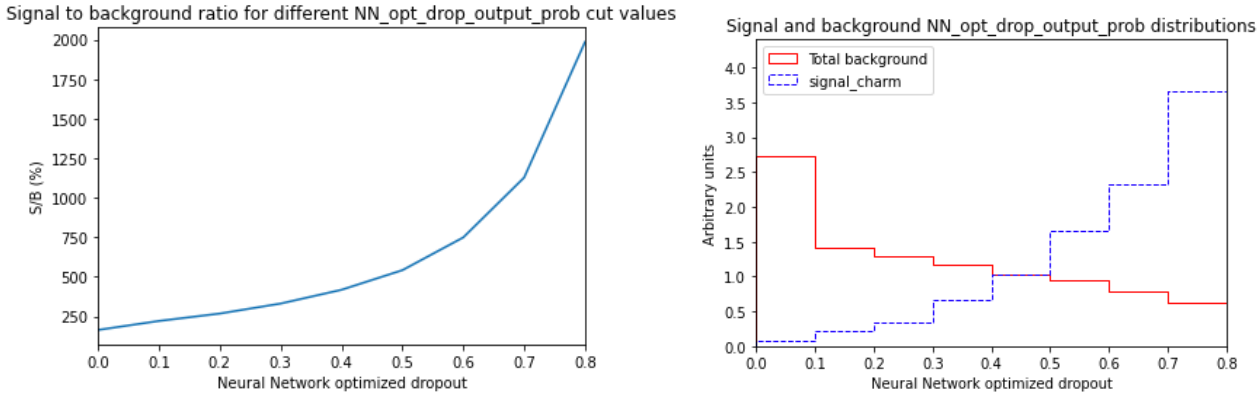
شکل ۱۰. مقادیر تفاوت loss (سمت چپ) و دقت accuracy (سمت راست) برای داده های آموزشی و تستی در هر دوره آموزش epoch. با تکرار روند آموزش مقدار پیش بینی شده توسط مدل به مقدار واقعی نزدیک شده و دقت افزایش می یابد.



شکل ۱۱. نسبت تعداد داده های سیگنال به پس زمینه (S/B) در سمت چپ و نمودار توزیع خروجی شبکه عصبی (احتمال سیگنال یا پس زمینه بودن) آنها در سمت راست.



شکل ۱۲. مقادیر تفاوت loss (سمت چپ) و دقت accuracy (سمت راست) برای داده های آموزشی و تستی در هر دوره آموزش epoch. با تکرار روند آموزش مقدار پیش بینی شده توسط مدل به مقدار واقعی نزدیک شده و دقت افزایش می یابد.

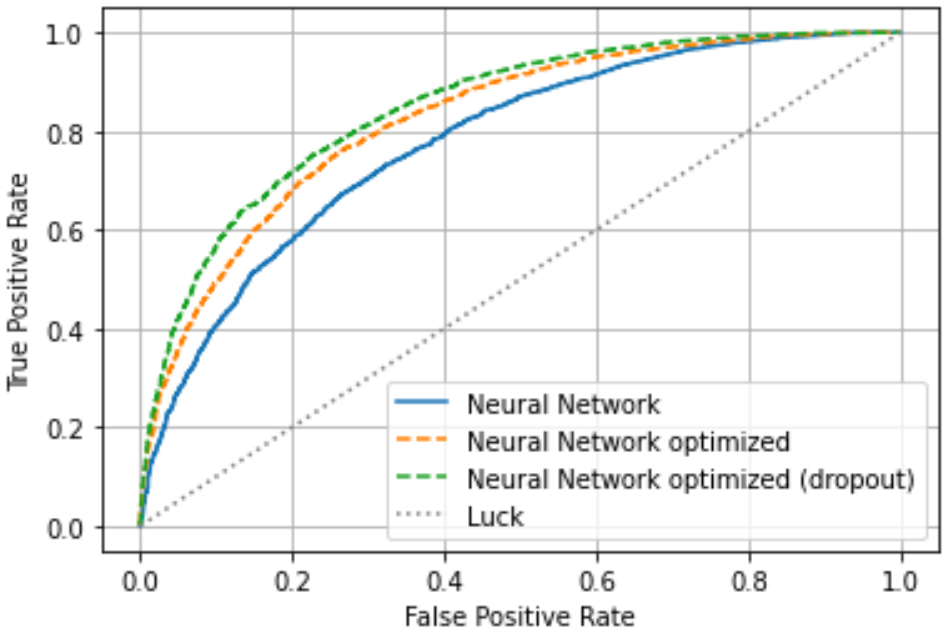


شکل ۱۳. نسبت تعداد داده های سیگنال به پس زمینه (S/B) در سمت چپ و نمودار توزیع خروجی شبکه عصبی (احتمال سیگنال یا پس زمینه بودن) آنها در سمت راست.

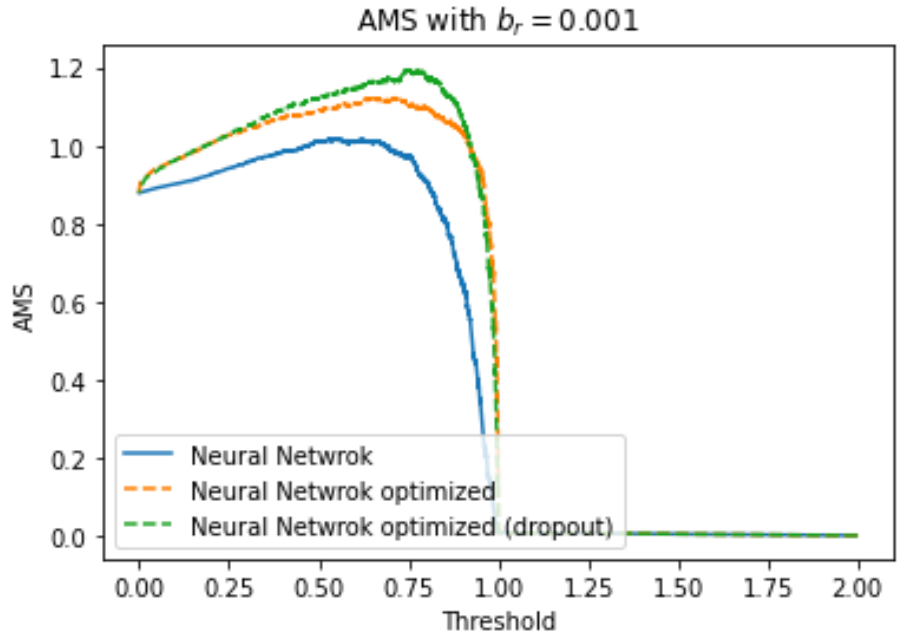
برای ارزیابی و تجسم عملکرد یک مدل طبقه بندی در یادگیری ماشین، منحنی مشخصه عملیاتی گیرنده (ROC) که یک نمایش گرافیکی است استفاده می‌شود. این منحنی نرخ مثبت واقعی TPR (سیگنالی که مدل به درستی سیگنال پیشبینی کرده) و نرخ مثبت کاذب FPR (پس زمینه ای که مدل به اشتباه سیگنال پیشبینی کرده) در آستانه های طبقه بندی مختلف را نشان می دهد. با تغییر آستانه ای که در آن پیش بینی ها به عنوان مثبت یا منفی طبقه بندی می شوند، می توانیم مقادیر مختلف TPR و FPR را محاسبه کنیم. منحنی ROC نشان می دهد که چگونه حساسیت و ویژگی مدل با حرکت آستانه تغییر می کند.

در حالت ایده آل، یک مدل طبقه بندی خوب باید TPR بالا و FPR پایین در مقادیر مختلف آستانه داشته باشد. منحنی ROC می تواند به تعیین آستانه بهینه برای مشکل خاص کمک کند، تعادل بین شناسایی مثبت های واقعی و اجتناب از مثبت های کاذب را متعادل کند. هرچه منحنی ROC به گوشه سمت چپ بالای نمودار نزدیکتر باشد، عملکرد مدل بهتر است. شکل ۱۴ بیانگر منحنی ROC برای مدل های شبکه عصبی استفاده شده در این آنالیز میباشد. با بهینه سازی پارامترهای مدل، شاهد عملکرد بهتر مدل ها در شناسایی داده های سیگنالی و پی زمینه ای میباشیم.

مقادیر مختلف آستانه (وزن مدل) منجر به باقی ماندن تعداد مختلف داده های سیگنالی و پس زمینه ای (همچنین نسبت سیگنال به پس زمینه) میشود. به منظوره تعیین ناحیه حساس به پدیده های سیگنالی در این آنالیز، آستانه با بالاترین مقدار AMS [۸] بعنوان برش استفاده شده است (شکل ۱۵). با در نظر گزفتن فرمول AMS، مقدار آستانه شامل بالاترین نسبت داده سیگنالی به پس زمینه ای میباشد.



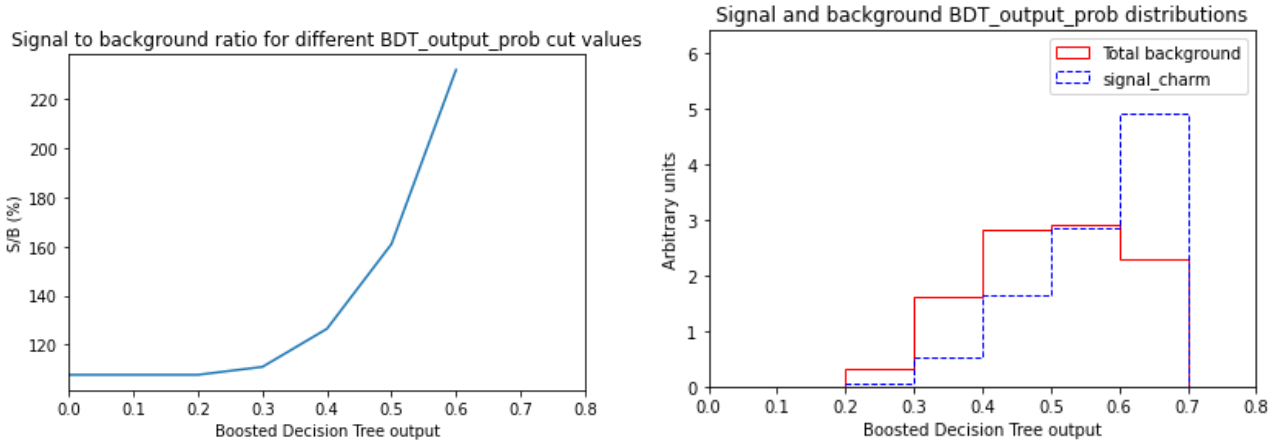
شکل ۱۴. منحنی ROC برای شبکه عصبی ساده (آبی)، شبکه عصبی بهینه شده (نارنجی) و شبکه عصبی بهینه شده به همراه لایه های حذفی (سبز).



شکل ۱۵. نمودار AMS برای شبکه عصبی ساده (آبی)، شبکه عصبی بهینه شده (نارنجی) و شبکه عصبی بهینه شده به همراه لایه های حذفی (سبز). آستانه با بالاترین مقدار AMS بهترین گزینه برای تعریف ناحیه سیگنالی میباشد.

- تصمیم درختی تقویت شده (Decision Tree)

درخت های تصمیم تقویت شده به عنوان الگوریتم های یادگیری ماشینی قدرتمند و پرکاربرد در زمینه علم داده و تجزیه و تحلیل ظاهر شده اند. با ترکیب نقاط قوت درخت های تصمیم گیری و تقویت الگوریتم ها، ثابت شده است که این مدل در حل مسائل پیچیده طبقه بندی و رگرسیون موثر میباشد. درخت‌های تصمیم، الگوریتم‌های یادگیری ماشینی تحت نظارت هستند که برای کارهای طبقه‌بندی و رگرسیون استفاده می‌شوند. آنها فضای ویژگی را به مناطق مختلف تقسیم می کنند، که توسط یک سری شرایط if-else هدایت می شوند. درختان تصمیم به دلیل قابلیت تفسیر و توانایی آنها برای مدیریت موثر هر دو ویژگی عددی و طبقه بندی کاربردی هستند. از سوی دیگر، الگوریتم‌های تقویتی، خانواده‌ای از الگوریتم‌های یادگیری ماشینی هستند که برای بهبود عملکرد مدل های ضعیف، با تمرکز هر مدل بر روی نمونه هایی که مدل های قبلی با آن ها مشکل داشتند میباشد. این فرآیند تکراری به حداقل رساندن خطاها و بهبود دقت کمک می کند. عملکرد مدل تصمیم درختی در جداسازی توزیع سیگنال و پس زمینه و نسبت داده های آنها در شکل ۱۶ نمایش داده شده است.

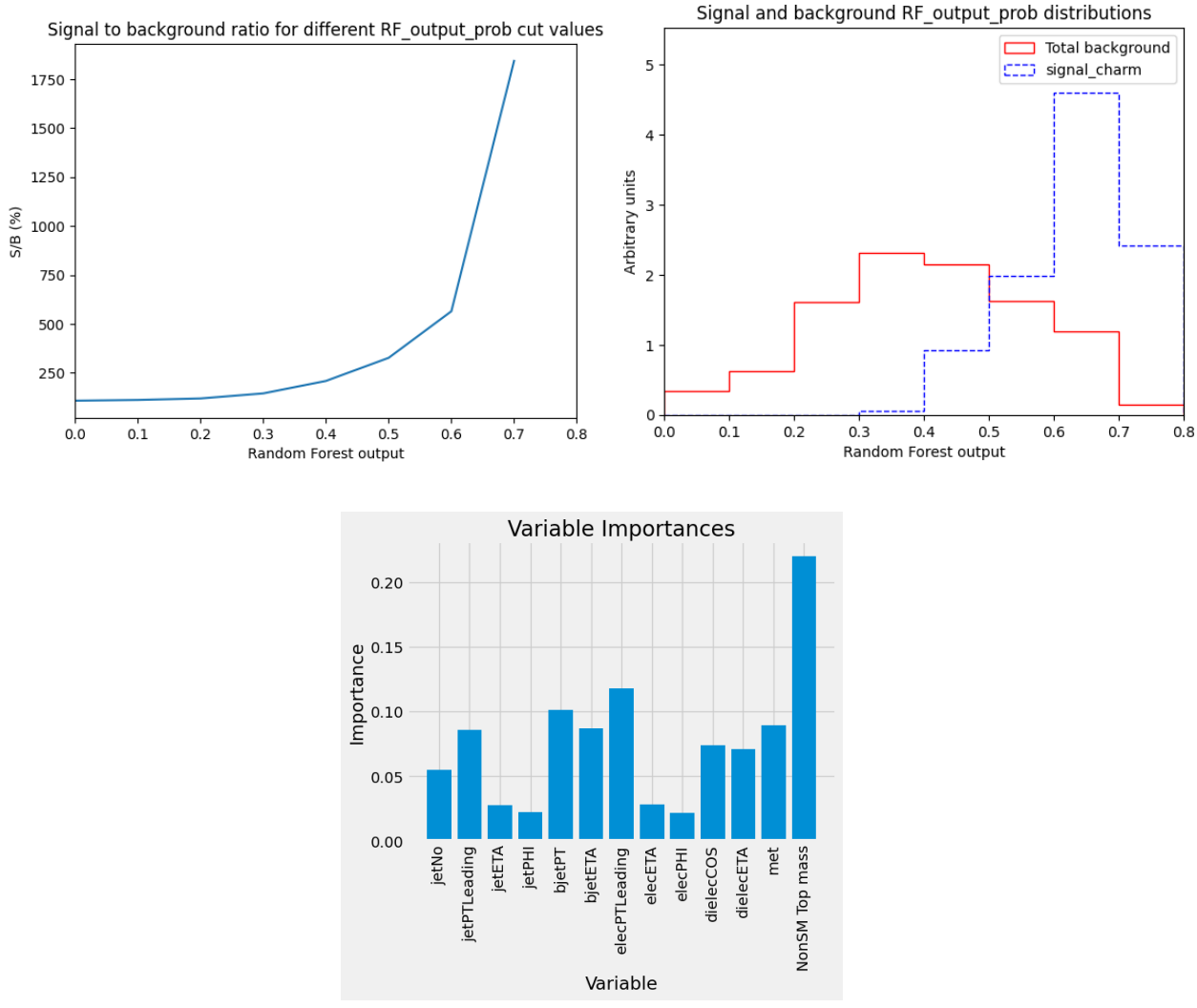


شکل ۱۶. نسبت تعداد داده های سیگنال به پس زمینه (S/B) در سمت چپ و نمودار توزیع خروجی تصمیم درختی (احتمال سیگنال یا پس زمینه بودن) آنها در سمت راست.

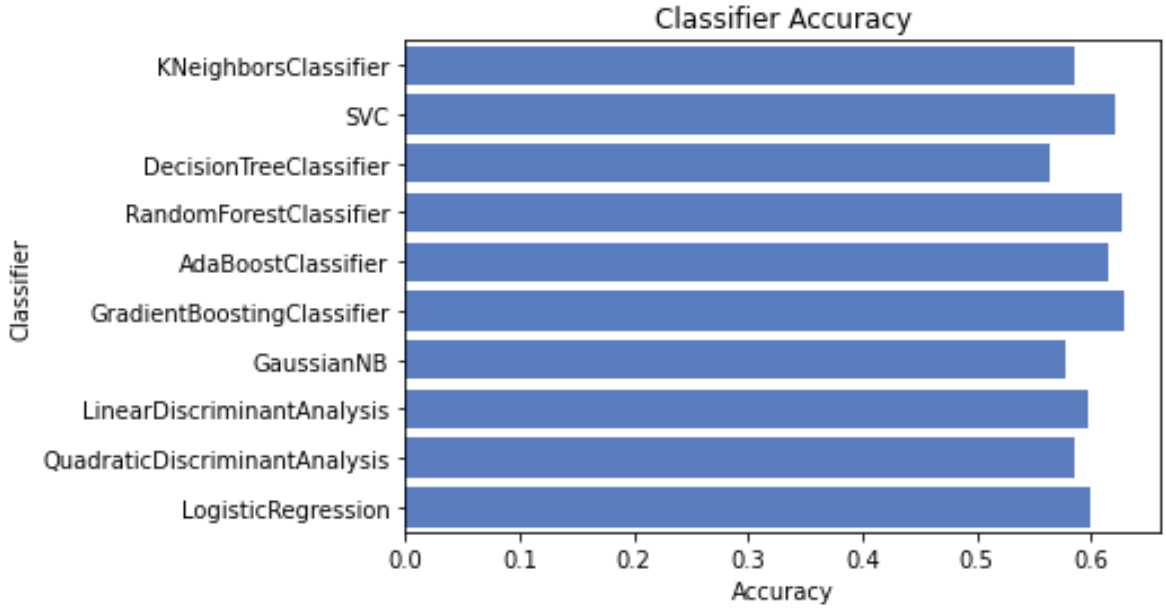
- جنگل تصادفی (Random Forest) و سایر مدل های طبقه بندی

در حوزه یادگیری ماشینی، Random Forest به عنوان یک الگوریتم برجسته است که قدرت یادگیری گروهی را با مفاهیم درخت تصمیم ترکیب می کند. جنگل تصادفی بر اساس اصل ترکیب درخت های تصمیم گیری چندگانه است. هر درخت تصمیم در مجموعه به طور مستقل بر روی زیر مجموعه‌های تصادفی داده‌های آموزشی رشد می‌کند و پیش‌بینی‌ها با تجمیع نتایج درختان جداگانه انجام می‌شود. هنگامی که درختان آموزش داده می شوند، با جمع آوری نتایج از هر درخت منفرد، پیش بینی ها انجام می شود. در کارهای طبقه بندی، کلاسی که اکثریت آرا را داشته باشد به عنوان پیش بینی نهایی انتخاب می شود، در حالی که در کارهای رگرسیونی، میانگین یا میانه مقادیر پیش بینی شده در نظر گرفته می شود.

عملکرد مدل تصمیم درختی در جداسازی توزیع سیگنال و پس زمینه، نسبت داده های آنها و اهمیت ویژگی های ورودی در شکل ۱۷ نمایش داده شده است. میزان دقت برای معروفترین مدل های طبقه بندی شده نیز در شکل ۱۸ بصورت نمودار میله نشان داده شده است.



شکل ۱۷. نسبت تعداد داده های سیگنال به پس زمینه (S/B) در سمت چپ و نمودار توزیع خروجی جنگل تصادفی (احتمال سیگنال یا پس زمینه بودن) آنها در سمت راست. اهمیت ویژگی های ورودی بصورت درصد در شکل پایین گزارش شده است.



شکل ۱۸. دقت بر حسب درصد برای مدل های طبقه بندی شده نظیر لاجیستیک خطی، بردار پشتیبان و گرادیان تقویت شده.

۵. مراحل بعدی این پژوهش:

بصورت خلاصه در این آنالیز، تمامی متغیرهای مستقل محاسبه شده برای پدیده های سیگنالی و پس زمینه ای بصورت ورودی به مدلهای هوش مصنوعی داده شده تا خروجی (وزن یا به معنای دیگر احتمال سیگنال یا پس زمینه بودن) مورد نظر بعنوان متغیر مستقل دیگر محاسبه شود. مدل های هوش مصنوعی زیادی برای موضوع طبقه بندی classification بین پدیده های سیگنالی یا پس زمینه ای قابل استفاده است، لذا سه عدد از آنها مورد مطالعه قرار گرفته شده است. مورد اول مدل تصمیم درختی Decision tree که یک مدل درخت مانند است که مجموعه ای از تصمیمات یا قوانین مورد استفاده برای طبقه بندی یا پیش بینی یک متغیر هدف را نشان می دهد. فرآیند انتخاب مهم ترین متغیر و تقسیم داده‌ها به صورت بازگشتی ادامه می یابد تا زمانی که زیرمجموعه‌های حاصل خالص باشند (یعنی فقط یک کلاس داشته باشند: سیگنال یا پس زمینه) یا تقسیم بیشتر دقت مدل را بهبود نمی‌بخشد. گره های نهایی درخت که برگ نامیده می شوند، کلاس یا مقدار پیش بینی شده را برای هر زیر مجموعه نشان می دهند. مورد دوم مدل شبکه عصبی Neural network شامل لایه‌هایی از گره‌های به هم پیوسته به نام نورون است که اطلاعات را پردازش و انتقال می‌دهد. داده‌های ورودی به لایه ورودی وارد می‌شوند و قبل از تولید خروجی، از طریق یک سری لایه‌های پنهان پردازش می‌شوند. هر نورون در شبکه ورودی از نورون‌های لایه قبلی دریافت می‌کند، تابع فعال‌سازی را روی مجموع وزنی ورودی‌ها اعمال می‌کند و نتیجه را به نورون‌های لایه بعدی ارسال می‌کند. شبکه‌های عصبی قادر به یادگیری روابط غیرخطی پیچیده بین ورودی‌ها و خروجی‌ها هستند و در طیف گسترده‌ای از کاربردها مانند بازسازی جت ها و لپتون ها، گیر اندازی ذرات در آشکارسازها و پردازش زبان طبیعی موفق بوده‌اند. مورد

آخر مدل جنگل تصادفی Random forest که یک روش یادگیری مجموعه ای برای طبقه بندی، رگرسیون و سایر موارد است که با ساختن تعداد زیادی درخت تصمیم به منظور آموزش عمل می‌کند. برای کارهای طبقه بندی، خروجی جنگل تصادفی کلاسی است که توسط اکثر درختان انتخاب شده است. در قدم بعدی وزن های محاسبه شده با مدل های هوش مصنوعی به همراه دیگر متغیرها ترکیب شده تا نواحی حساس به سیگنال signal regions مشخص شوند. به منظور انجام کارهای آماری میتوان از کتابخانه های تست آماری نظیر pyhf [۹] استفاده کرد و بعنوان مثال اهمیت significance پدیده های سیگنالی نسبت به پس زمینه ای را در نواحی حساس محاسبه کرد. مقادیر مختلف عدم قطعیت uncertainty مانند ۵۰٪ یا ۱۰۰٪ به همراه تعداد داده های سیگنال و پس زمینه بعنوان ورودی به تابع BinomialExpZ داده شده تا مقادیر اهمیت برای مقیاس های فیزیک جدید Λ محاسبه شود. در انتها با استفاده از آنالیزهای آماری حد بالا برای مقیاس فیزیک جدید Λ با ضریب ۹۵٪ سطح اطمینان برای سیگنال های tt و tW گزارش داده خواهد شد.

فهرست منابع:

1. Alan Axelrod, Flavor Changing Z0 Decay and the Top Quark, Volume 209, Issue 2, 27 December (1982).
2. Xue-Qian Li, et.al, The Production of t anti-c or anti-t c quark pair by e^+ e^- collision based on the standard model and its extensions, Physics Letters B Volume 313, Issues 3–4, 2 September (1993).
3. S. Weinberg, A Model of Leptons, Phys. Rev. Lettr. 19, 1264 November (1967).
4. The CMS Collaboration, CMS Physics Analysis Summary, 7 November (2017)
5. Ian J Goodfellow, Jean Pouget, Peter Surberg, Generative Adversarial Network, arXiv:1406.2661 (2014).
6. Nicolas Kriegestroke, et.al, Neural network models and deep learning-a primer for biologists, arXiv:1902..47.4 (2019).
7. Ryan Nash, Kiern O'shea, Introduction to convolutional neural network, arXiv:1511.8458 (2013).
8. Breiman, L. Random Forest-Machine learning, springer 2001, DOI https://doi.org/10.1023/A:1010933404324.
9. Giordon Stark, et.al, pure-Python implementation of HistFactory with tensors and automatic differentiation, arXiv:2211.15838v1 [hep-ex] (2022).