

**A Stacking-Based Model for Non-Invasive Detection of Coronary Heart Disease**

Coronary arteriongraphy (CAG) is an accurate invasive technique for the diagnosis of coronary heart disease (CHD).

Arash Mehrzadi, Hamidreza Hematyar

QIAU

**مقدمه**

عروق کرونر (CAG) یک تکنیک تهاجمی دقیق برای تشخیص بیماری عروق کرونر قلب (CHD) است. با این حال، روش تهاجمی آن برای تشخیص CHD در معاینه فیزیکی سالانه مناسب نیست.

با کاربرد موفقیت‌آمیز یادگیری ماشین (ML) در زمینه‌های مختلف، هدف ما انجام یکپارچه‌سازی انتخابی الگوریتم‌های چندگانه ML و تأیید اعتبار روش‌های انتخاب ویژگی با اطلاعات بالینی شخصی است که معمولاً در معاینه فیزیکی سالانه دیده می‌شود.

در این مطالعه، یک مدل مبتنی بر انباشته دو سطحی طراحی شده است که در آن سطح 1 سطح پایه و سطح 2 متا سطح است. پیش‌بینی طبقه‌بندی‌کننده‌های سطح پایه به عنوان ورودی فراسطح انتخاب می‌شوند. ابتدا ضریب همبستگی پیرسون و حداکثر ضریب اطلاعات محاسبه می شود تا طبقه بندی کننده با کمترین همبستگی پیدا شود. سپس از الگوریتم شمارش برای یافتن بهترین طبقه‌بندی‌کننده‌های ترکیبی استفاده می‌شود که در پایان بهترین نتیجه را به دست می‌آورند. مجموعه داده Z-Alizadeh Sani CHD که ما استفاده می کنیم شامل 303 مورد است که توسط CAG تأیید شده است.

نتایج تجربی نشان می دهد که مدل پیشنهادی، دقت، حساسیت و ویژگی به ترتیب 95.43٪، 95.84٪، 94.44٪ برای تشخیص CHD به دست می آورد. روش پیشنهادی می‌تواند به طور موثر به پزشکان کمک کند تا افراد دارای عروق کرونر طبیعی از مبتلایان به CHD را تشخیص دهند.

بیماری عروق کرونر قلب (CHD) یکی از علل اصلی مرگ و میر قلبی عروقی در سطح جهان است. در حال حاضر روش های تشخیصی CHD را می توان به روش های تهاجمی و غیرتهاجمی تقسیم کرد. آنژیوگرافی عروق کرونر (CAG) یک تکنیک تشخیصی تهاجمی نسبتاً ایمن و قابل اعتماد است که به طور گسترده در عمل بالینی به عنوان استاندارد طلایی برای تشخیص CHD مورد استفاده قرار گرفته است. با این حال، ماهیت تهاجمی آن و هزینه عملیات نسبتاً گران، استفاده از آن در معاینه فیزیکی سالانه را دشوار می کند. الکتروکاردیوگرام (ECG) و اکوکاردیوگرافی روش‌های غیرتهاجمی هستند، اما دقت قابل اعتمادی ندارند.

بنابراین، یافتن روش‌های غیرتهاجمی جدید برای تشخیص CHD ضروری است.

در قلب و عروق بالینی، یادگیری ماشین (ML) یک روش موثر برای پیش‌بینی مرگ و میر ناشی از همه علل در بیماران مشکوک به CHD ثابت شده است. در اپیدمیولوژی قلبی عروقی تحت بالینی، ML می تواند پیش بینی بهتری نسبت به نمرات استاندارد خطر قلبی عروقی در ارتباط با نقاط داده فنوتیپی ارائه دهد. روش های ML به طور گسترده در برخورد با داده های موجود در پزشکی استفاده می شود. در سال‌های اخیر، تعدادی از الگوریتم‌های ML برای تشخیص CHD توسعه یافته‌اند. Feshki و Shijani تشخیص CHD را با یک الگوریتم تکاملی و یک شبکه عصبی پیشخور بهبود دادند. Davari و همکاران ویژگی های ECG را با روش های فرکانس و دامنه غیرخطی برای شناسایی علائم CHD با طبقه بندی کننده بردار پشتیبان (SVC) استخراج کرد.

Vernekar و همکاران ویژگی های مارکوف را به همراه سایر ویژگی های حوزه آماری و فرکانس از فونوکاردیوگرام (PCG) استخراج کرد و از مجموعه شبکه عصبی مصنوعی و درخت افزایش گرادیان برای آموزش مدل استفاده کرد.

Kumar و همکاران همچنین از سیگنال های ECG اما با تبدیل موجک تحلیلی انعطاف پذیر برای مشخص کردن CHD استفاده کرد.

یک روش ترکیبی پیشنهاد کرد که شامل شناسایی عوامل خطر با استفاده از انتخاب زیر مجموعه ویژگی مبتنی بر همبستگی با روش جستجوی بهینه‌سازی شنای ذرات و الگوریتم‌های خوشه‌بندی K-means بود [9]. علیزاده و همکاران از سه طبقه‌بندی کننده برای تشخیص تنگی سه شریان کرونری، یعنی نزولی قدامی چپ، سیرکومفلکس چپ و شریان کرونری راست استفاده کرد تا دقت بالاتری برای تشخیص CHD بدست آورد.

داوری و همکاران با پایگاه داده Long Term ST به دقت تشخیص 99.2% دست یافتند، اما پایگاه داده ای که آنها برای بیماران CHD استفاده کردند با تغییرات بخش ST مختلف همراه بود.

و در عمل بالینی، بسیاری از بیماران CHD قطعه ST طبیعی دارند. بنابراین، استفاده از پایگاه‌های اطلاعاتی بیماران مبتلا به بیماری عروق کرونر اما بخش‌های ST طبیعی ممکن است برای استفاده از مدل تشخیص CHD مبتنی بر هوش مصنوعی در موقعیت‌های پیچیده بالینی مفیدتر باشد. علاوه بر این، تحقیقات قبلی معمولاً تنها از یک نوع طبقه‌بندی کننده ML برای تشخیص خودکار CHD استفاده می‌کردند. با این حال، بسیاری از محققان ML به ویژه آنهایی که در مسابقات ML شرکت می کنند، با موفقیت از تکنیک های ترکیب طبقه بندی کننده برای بهبود دقت طبقه بندی کننده ها استفاده کرده اند .

تکنیک‌های ترکیب پیش‌بینی‌های به‌دست‌آمده از طبقه‌بندی‌کننده‌های چندگانه سطح پایه را می‌توان در سه چارچوب ترکیبی خلاصه کرد: رأی‌گیری (مورد استفاده در بسته‌بندی و تقویت)، پشته‌بندی و آبشاری. برای مجموعه داده های پیچیده تر، طبقه بندی کننده سنتی را می توان با انواع مختلفی از قوانین ترکیبی بهبود بخشید . در پشته‌بندی، پیش‌بینی‌های مجموعه‌ای از طبقه‌بندی‌کننده‌ها به عنوان ورودی‌های الگوریتم یادگیری سطح بعدی ارائه می‌شوند. سطح بعدی الگوریتم برای ارتباط بهینه پیش بینی های مدل و تشکیل سطح بعدی مجموعه نهایی پیش بینی ها آموزش داده شده است. روابط جفتی همیشه بین سطوح مختلف قبل از پیش بینی نهایی وجود دارد. ما روابط بین مدل‌ها را در سطح پایه تحلیل می‌کنیم و ترکیب بهینه مدل را با یک الگوریتم شمارش پیدا می‌کنیم.

به طور خلاصه، سهم اصلی این کار به شرح زیر است:

هشت روش انتخاب ویژگی برای ارزیابی عملکرد آنها برای تشخیص خودکار CHD بررسی شده است. ما متوجه شدیم که استراتژی یادگیری ماشینی RFECV بالاترین عملکرد پیش‌بینی را در اعتبارسنجی متقابل ده برابری مکرر به دست آورد. آن ویژگی‌هایی که با روش RFECV انتخاب می‌شوند برای متخصصان قلب در تشخیص بالینی CHD ارزش مرجع بالایی دارند.

در مجموع از 10 روش طبقه بندی استفاده می شود. با تجزیه و تحلیل نتایج، مشخص شد که ترکیب مدلی که بهترین عملکرد را نشان می‌دهد را نمی‌توان با محاسبه مستقیم ضریب همبستگی پیرسون (PCC) و حداکثر ضریب اطلاعات (MIC) تعیین کرد. بنابراین، یک استراتژی جدید برای جستجوی ترکیب بهینه پیشنهاد می‌شود که در آن ابتدا مدلی انتخاب می‌شود که حداقل همبستگی را با سایر مدل‌ها داشته باشد و سپس با برشمردن هر ترکیب احتمالی مدل انتخاب‌شده با سایر مدل‌ها، ترکیب بهینه تعیین می‌شود. نتایج ما نشان می دهد که استراتژی پیشنهادی عملکرد رضایت بخشی دارد.

ترکیب مدل بهینه برای تشخیص خودکار CHD تعیین می شود. استفاده از استراتژی ترکیب مدل پیشنهادی در 3 مجموعه داده دیگر نیز نتایج رضایت‌بخشی را نشان می‌دهد، که توانایی تعمیم استراتژی ترکیب مدل پیشنهادی ما را نشان می‌دهد.

ادامه این مقاله به شرح زیر سازماندهی شده است. در بخش دوم، منبع داده و روش های پیش پردازش داده ها معرفی شده است. در بخش III، جزئیات فنی مدل مبتنی بر انباشته دو سطحی پیشنهادی ما شرح داده شده است. نتایج تجربی در بخش IV و سپس بحث در بخش V ارائه شده است.

**دیتا**

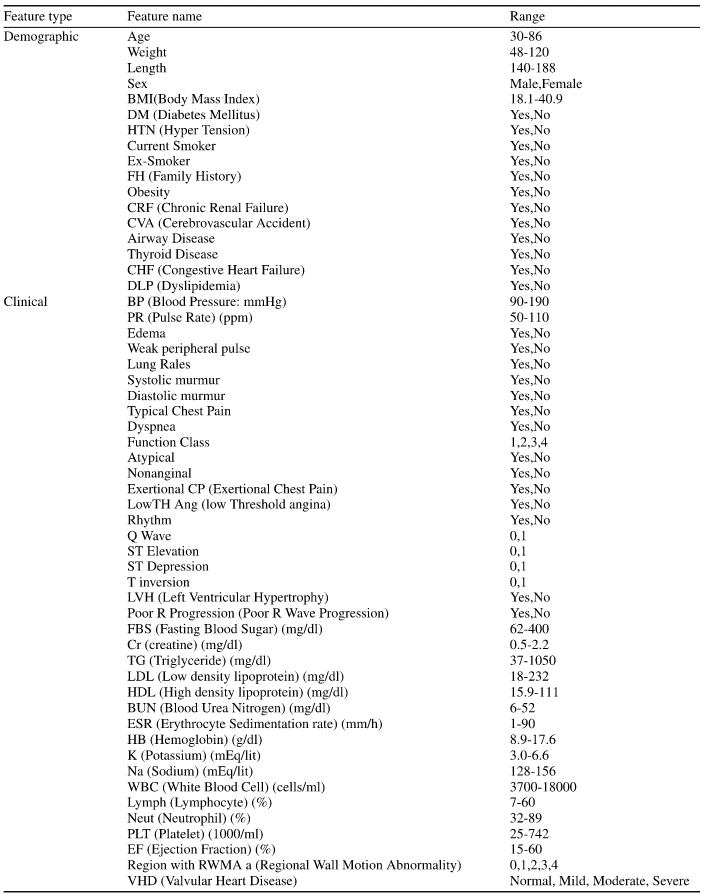
مجموعه داده Z-Alizadeh شامل 216 بیمار CHD و 87 فرد سالم است که توسط 54 نوع مختلف نشان داده شده اند. ویژگی های بالینی و جمعیت شناختی همانطور که در جدول زیر نشان داده شده است.

مجموعه داده‌ها عدم تعادل بزرگی را در توزیع طبقات هدف نشان می‌دهد، زیرا تقریباً 3 برابر بیشتر از افراد سالم بیماران CHD وجود دارد. در چنین حالتی، روش نمونه‌برداری بیش از حد اقلیت مصنوعی (SMOTE) برای حل مشکل عدم تعادل استفاده می‌شود. ایده اصلی روش SMOTE تجزیه و تحلیل کلاس های اقلیت و ترکیب کلاس های اقلیت جدید با نمونه برداری بیش از حد است.

داده های افراد عادی توسط SMOTE در طول اعتبارسنجی متقاطع و نه قبل از فرآیند اعتبار سنجی متقابل نمونه برداری می شود. داده های مصنوعی فقط برای مجموعه آموزشی ایجاد می شود بدون اینکه بر مجموعه آزمایشی تأثیر بگذارد. اگر یک ویژگی دارای واریانسی باشد که مرتبه‌ای بزرگتر از سایر ویژگی‌ها باشد، ممکن است بر تابع هدف تأثیر بگذارد و باعث می‌شود برآوردگر نتواند همانطور که انتظار می‌رود از ویژگی‌های دیگر به درستی یاد بگیرد.

از آنجایی که 54 ویژگی مجموعه داده شامل 23 داده عددی و 31 داده طبقه‌بندی می‌شود، از تکنیک حداکثر و حداقل نرمال‌سازی برای استانداردسازی این ویژگی‌ها استفاده می‌شود. حداکثر و حداقل نرمال سازی یک روش رایج پردازش داده ها است که می توان آن را به صورت (1) تعریف کرد. x ویژگی ورودی، max نشان دهنده حداکثر مقدار، min نشان دهنده حداقل مقدار، و x∗ نشان دهنده مقدار خروجی پس از عادی سازی است. در این مطالعه، ما از این رویکرد برای مقیاس‌بندی 23 ویژگی استفاده می‌کنیم. یافتن روابط اهمیت بالقوه در میان ویژگی ها مفید است.

در ادامه، فیچر های دیتاست ذکر شده است:



**پیش پردازش داده ها**

با توجه به تنوع داده ها در انواع مختلف، در ابتدا نیاز داریم دیتای مورد نظر را به فرمی برسانیم که برای الگوریتم های ما قابل پذیرش و قابل پردازش میباشد.

دیتاست مورد بحث دارای انواع زیر میباشد:

* Integer
* Bool
* Enum

که برای هر یک از این داده ها نیاز به رویکردی متفاوت داریم، و برای این کار دیتاست را به سه بخش متفاوت تقیسم میکنیم تا بتوانیم فرایند های مربوط به هر کدام از این تایپ را انجام دهیم و سپس دیتاست ها را به یکدیگر چسبانده و دیتافریمی یکپارچه تهیه میکنیم.

برای دیتاهایی با تایپ عددی از رویکرد MinMaxScaler برای اسکیل کردن مقادیر بین دو عدد ۰ و ۱ استفاده میکنیم.

همچنین برای دیتاهایی با تایپ ‌bool از رویکرد binary استفاده میکنیم که مقادیر صحیح به مقدار ۱ و مقادیر نادرست به مقدار ۰ مپ میشوند

در نهایت برای مقادیر Enum نیز از تکنیک one hot encode استفاده میکنیم.

سپس با چسباندن بخش های جدا شده به یکدیگر، به دیتابیس واحدی میرسیم که قابل ارایه به مدل های تعریف شده میباشد.

فیچر سلکشن

انتخاب ویژگی ها در برخورد با ویژگی های اضافی اهمیت زیادی دارد. سه معیار رایج انتخاب ویژگی شامل فیلتر، لفاف و تعبیه شده است.

روش های فیلتر، رابطه بین ویژگی ها و برچسب را با استفاده از ابزارهای آماری شامل واریانس، اطلاعات متقابل و آزمون مجذور کای (CHI2) محاسبه می کنند. روش های wrapper ارتباط نزدیکی با طبقه بندی کننده دارند. اصل روش wrapper انتخاب بهترین زیرمجموعه با توجه به عملکرد طبقه بندی کننده است.

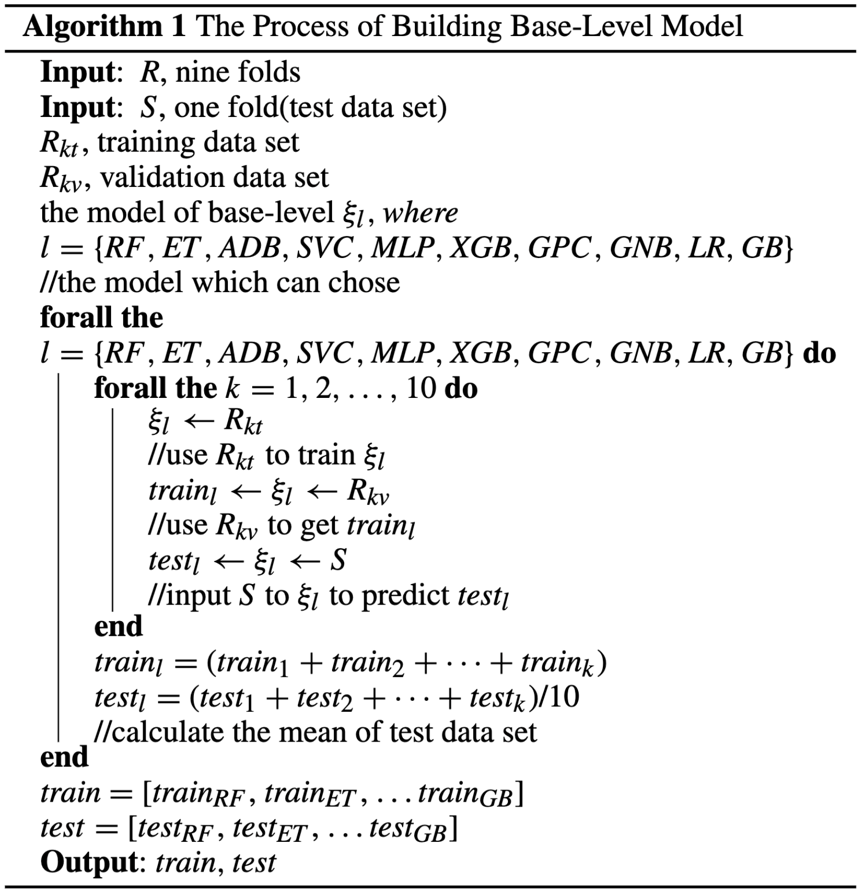
علاوه بر این، حذف ویژگی بازگشتی با اعتبارسنجی متقاطع (RFECV) می تواند تأثیر تنظیم مصنوعی تعداد ویژگی های باقی مانده در مجموعه ویژگی را از بین ببرد. روش های تعبیه شده با فرآیند آموزش مدل ادغام می شوند تا ویژگی ها به طور خودکار انتخاب شوند. افزایش گرادیان شدید (XGB) به دلیل کارایی بالا به طور گسترده ای به عنوان یک روش انتخاب ویژگی تعبیه شده استفاده شده است [26].

مدل مقاله

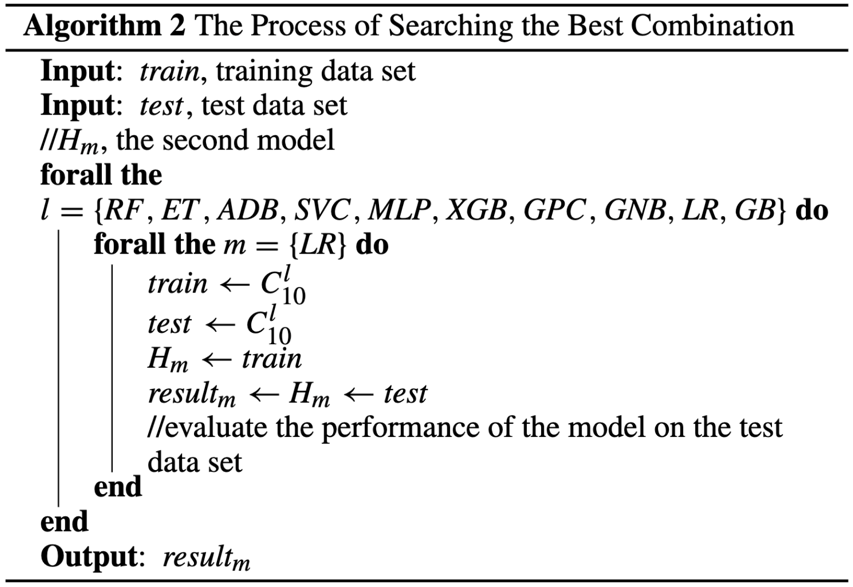
مدل پیشنهادی عمدتاً از دو سطح تشکیل شده است که در آن سطح 1 سطح پایه و سطح 2 متا سطح است. پیش‌بینی طبقه‌بندی‌کننده‌های سطح پایه به عنوان ورودی فراسطح انتخاب می‌شوند. سطح پایه شامل 10 مدل از scikit-learn، از جمله جنگل تصادفی (RF)، درختان اضافی (ET)، adaBoost (ADB)، SVC، پرسپترون چند لایه (MLP)، XGB، طبقه‌بندی فرآیند گاوسی (GPC)، گاوسی خلیج ساده (GNB)، رگرسیون لجستیک (LR)، تقویت گرادیان (GB) [27]-[36]. عملکرد طرح‌های انباشتگی تحت تأثیر تعداد طبقه‌بندی‌کننده‌های سطح پایه است [37]. به طور کلی، طبقه‌بندی‌کننده‌های سطح پایه با پیش‌بینی‌های همبستگی ضعیف عملکرد خوبی دارند [37]. PCC و MIC را می توان به عنوان معیاری برای تعیین کمیت ارتباط و افزونگی در بین ویژگی ها استفاده کرد [38]-[40]، با نزدیک تر بودن به 0 نشان دهنده همبستگی ضعیف تر است. سپس، از الگوریتم شمارش برای جستجوی بهترین طبقه‌بندی‌کننده‌های ترکیبی استفاده می‌کنیم.

ما دو الگوریتم را خلاصه می‌کنیم که می‌توانند فرآیند انباشته‌سازی و شمارش را نشان دهند. مجموعه داده ابتدا به صورت تصادفی مخلوط شده و به 10 برابر تقسیم می شود. برای هر فولد، یک فولد به عنوان داده آزمایشی (S) و چین های باقیمانده به صورت R در نظر گرفته می شود. کل فرآیند 10 بار تکرار می شود. R و S ورودی Alg.1 هستند. Alg.1 عمدتا شامل دو حلقه است. حلقه اول فرآیند ساخت ده مدل سطح پایه است و حلقه دوم فرآیند اعتبارسنجی متقاطع 10 برابری برای تولید داده های آموزشی و آزمایشی است. R نیز به 10 تا تقسیم می شوند. یک فولد به عنوان مجموعه اعتبارسنجی (Rkv) و تاهای باقیمانده به عنوان مجموعه داده های آموزشی (Rkt) در نظر گرفته می شود. Rkt وارد مدل سطح پایه مورد استفاده برای آموزش مدل سطح پایه (ξl) می شود.

Rkv برای تولید قطار استفاده می شود. بعداً S برای تولید testl وارد مدل سطح پایه ξl می شود. از آنجایی که حلقه 10 بار تکرار می شود، Trainl دقیقا برابر با مجموع ده برابر است و مجموعه داده های تست باید میانگین شود. در نهایت، مجموعه آموزش و آزمون اتحادیه تولید شده توسط 10 مدل پایه مختلف به عنوان خروجی گرفته می شود.

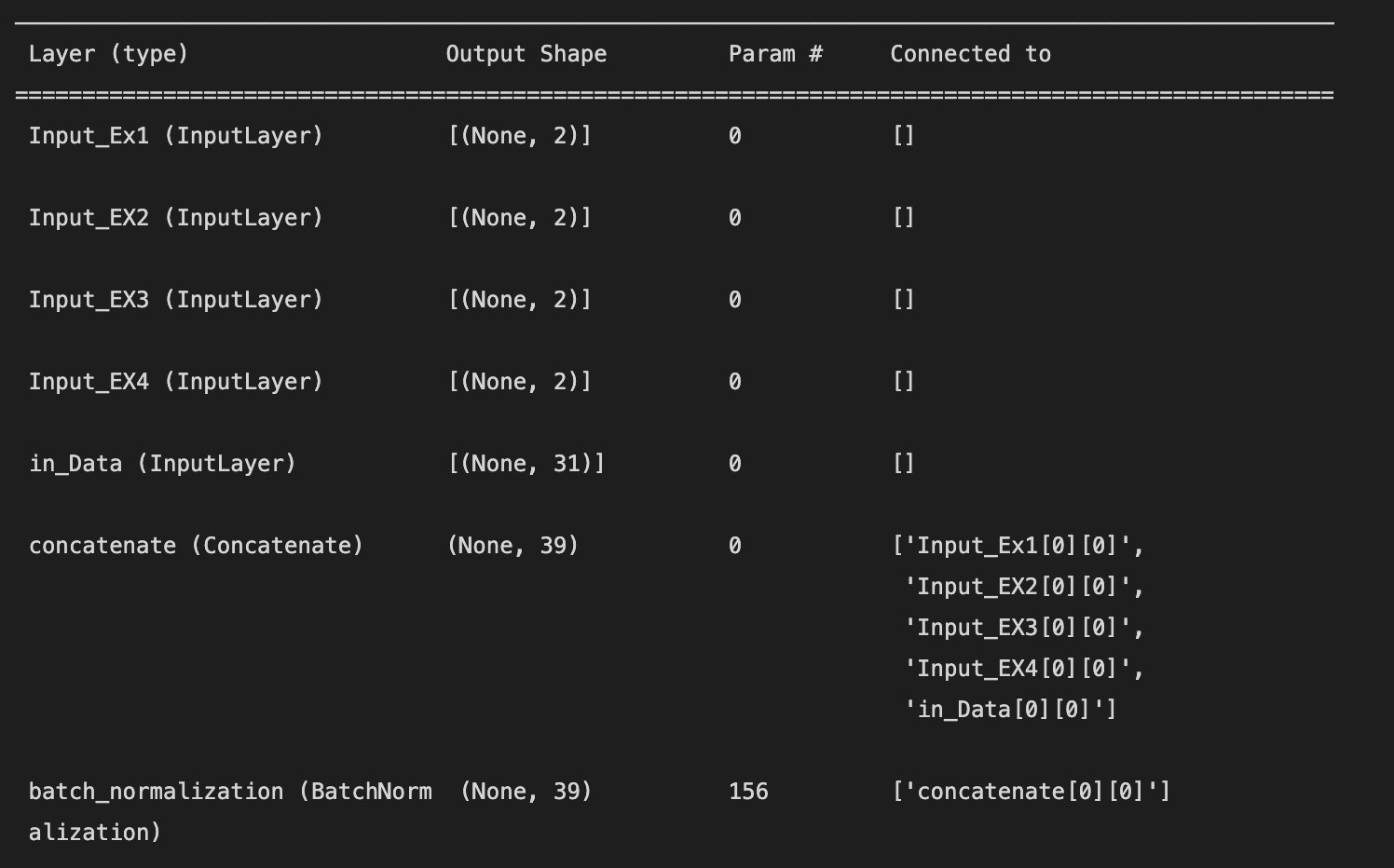


خروجی Alg. 1 به عنوان ویژگی های جدید سطح متا در نظر گرفته می شود. از آنجایی که استفاده مستقیم از همه ویژگی های جدید بدون فیلتر کردن عاقلانه نیست، Alg. 2 برای جستجوی ترکیب بهینه استفاده می شود. Alg. 2 عمدتا شامل دو حلقه است. در حلقه اول، 10 نوع ترکیب ممکن وجود دارد، از جمله C110، C210، C310، C410، C510، C610، C710، C810، C910، C1010 به عنوان ورودی حلقه دوم. همه ترکیبات ممکن (بدون تکرار آنها) به جای قرار دادن همه آنها در حلقه بعدی، شمارش می شوند. در حلقه دوم از ورودی قطار برای آموزش مدل Hm استفاده می شود. مدل LR برای کاهش پیچیدگی مدل استفاده می شود [37]. سپس آزمون به مدل آموزش دیده (Hm) وارد می شود تا عملکرد مدل در مجموعه داده های آزمون ارزیابی شود. در نهایت ترکیب مدل با بالاترین دقت تعیین می شود.



**مدل پیشنهادی**

مدل پیشنهادی ما شامل متدهای rf mlp lr و et است. مادر مدل پیشنهادی خود از روش mixture of expe استفاده کرده ایم دلیل استفاده ما از این روش این است که متد های اولیه مورد ارزیابی در مقاله اخیر می توانند فرضیه های متفاوتی را در فضای نمونه ای مثال های ما ایجاد کند که هرکدام منحصر به فرد و دارای ویژگی های خاص خود است. با استفاده از این روش می توانیم از تمام ویژگی های فرضیه های ایجاد شده توسط بهترین متد های موجود استفاده کنیم و از تجمیع و پردازش آنها به نتایج بهتری برسیم بدنه اصلی مدل پیشنهادی ما شامل لایه های dense variential به همراه تابع prior و posterior است که امکان ایجاد مدل های bayesian را فراهم میکند. ما از ۲ لایه dense variential استفاده کرده ایم و خروجی متدهای انتخاب شده با داده های خروجی rfecv تجمیع شده و به عنوان ورودی مدل bayesian استفاده شده است ( متدولوژی پیشنهادی با انواع داده ها سازگار بوده و مبتنی بر نوع ورودی متد های اولیه را آموزش می دهد و خروجی آنها را با همان داده های اولیه تجمیع کرده و به لایه های dense variential انتقال داده میشوند )



مقایسه نتایج

جدول دوم

مقایسه بازده پیش پردازش دیتا



جدول سوم

نتایج انتخاب ویژگی با استفاده از متد های مختلف





جدول چهارم

دقت تشخیص CHD برای روش های مختلف انتخاب ویژگی با مقادیر k مختلف.



جدول پنجم

اجرای روش های مختلف تست با روش انتخاب ویژگی RFECV و XGB.



جدول ششم

مقایسه عملکرد روش های مختلف تست



متد ها



بررسی متد ها