باسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی برق

درس هوش محاسباتی

فاز سوم پروژه درس

علی محرابیان96102331

استاد: دکتر رضایی

تابستان 1399

در فاز های قبلی به کمک سیستم های فازی و شبکه های عصبی،سعی در تشخیص بیماری های قلبی داشتیم.یکی از مراحل مهمی که در فاز 1 به طور کامل به آن پرداختیم،بحث انتخاب ویژگی های موثر هنگام آموزش داده ها است.ابتدا خلاصه ای از روشی که در فاز 1 به کار بردیم را بیان می کنیم،سپس انتخاب ویژگی های موثر را به کمک الگوریتم ژنتیک در این فاز انجام می دهیم.

بخش اول:مروری دوباره بر feature selection

از کارهای مرسوم در مقاله های learning،بحث feature selection و pre-process ماتریس ویژگی است.
در این پروژه از الگوریتم ReliefF برای انتخاب ویژگی استفاده کردیم .در این جا شرح مختصری از آن را انجام می دهیم.

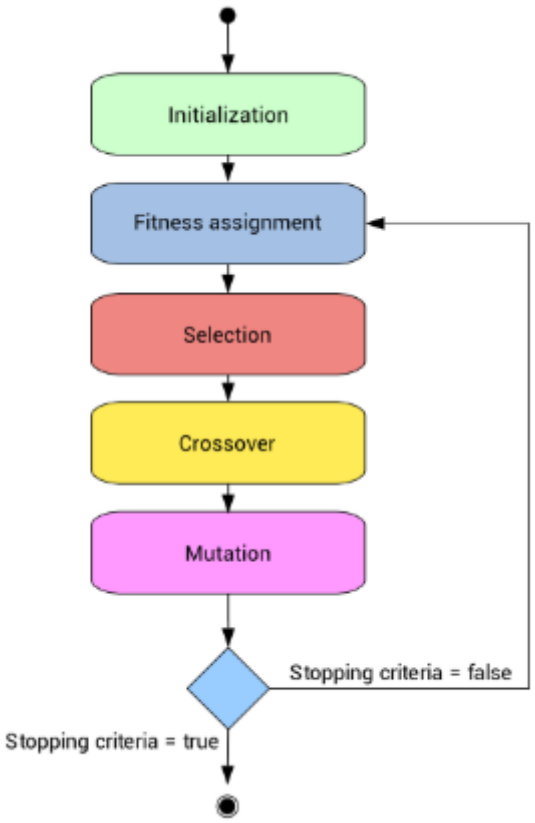
در این روش برای هر ویژگی در ابتدا یک وزن 0 در نظر گرفته می شود.در هر بار تکرار،یک نمونه به طور تصادفی انتخاب می شود و به کمک الگوریتم KNN،بردار های ویژگی از تک تک کلاس ها که نزدیک به نمونه انتخاب شده هستند را پیدا می کنیم.اگر هر دو از یک کلاس بودند،امتیاز آن کاهش و اگر از کلاس متفاوت بودند،امتیاز آن افزایش می یابد.امیتاز هم به صورت فاصله اقلیدسی تعریف می شود.

$$w_i = w_i - (x_i - nearheat_i)^2 + (x_i - nearmiss_i)^2$$

به کمک کران چبی شف ثابت می شود که می توان یک upperbound خوب برای خطا پیدا کرد.

بخش دوم:انتخاب ویژگی های موثر به کمک الگوریتم ژنتیک:

فرآیند انتخاب ویژگی های موثر از دید ریاضی به عنوان یک مسئله بهینه سازی مطرح می شود.یکی از راه حل های مفید برای حل این مسئله،استفاده از الگوریتم ژنتیک است.در ابتدا یک شمای کلی از مراحلی که در پیاده سازی الگوریتم های ژنتیک با آن سرو کار داریم را در تصویر زیر می بینیم.



حال هر قسمت را به طور کامل در صفحات بعد شرح می دهیم.

2.1) initialization:

از آن جایی که الگوریتم ژنتیک یک روش آماری و تصادفی برای بهینه سازی است،لذا به طور تصادفی به تولید جمعیت اولیه می پردازیم.در هر مرحله جمعیت را برابر با 4 individual در نظر می گیریم.هر individual از 30 ژن تشکیل می شود.روش رمز گذاری به این صورت است که در هر مرحله،ویژگی مربوط به هر ژن که در فرآیند آموزش حضور دارد،برابر با ژن 1 و عدم حضور ویژگی برابر با ژن 0 است.همان طور که اشاره کردیم،برای تولید جمعیت اولیه،رشته باینری 0 و 1 را به صورت تصادفی تولید می کنیم.

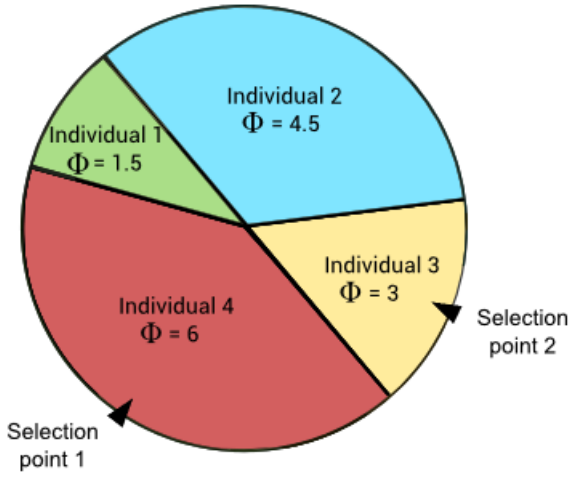
2.2) fitness assignment:

مرحله بعدی،تخصیص یک مقدار fitness به هر individual در جمعیت می باشد.در این مرحله از شبکه عصبی RBF که در فازهای قبل،بهترین خروجی را برای ما داشت استفاده می کنیم.در هر مرحله،برای هر individual، داده های خود را به دو دسته train و test تقسیم کرده و به کمک ویژگی های حاضر،شبکه عصبی را آموزش داده و در نهایت به کمک داده های تست،نسبت بیمارانی که برچسب غلط می خورند به کل بیماران را به عنوان مقدار متناسب با هر individual در نظر می گیریم.پس در هر مرحله،4 شبکه عصبی متفاوت آموزش داده می شود و در نتیجه 4 مقدار به دست آمده متناسب با هر individual خواهیم داشت.جزئیات پیاده سازی شبکه عصبی RBF در فاز قبلی به طورکامل شرح داده شده است.

حال به کمک روش ranked-based،ابتدا مقادیر به دست آمده را به صورت نزولی sort کرده،در نتیجه Individual ها rank بندی شده و به کمک یک ضریب ثابت مانند k، k برابر rank هر individual را برابر با مقدار fitness آن در نظر می گیریم.هر چه مقدار ضریب ثابت بیشتر باشد،احتمال انتخاب individual با خطای کمتر ،بیشتر می باشد.

selection(2.3:

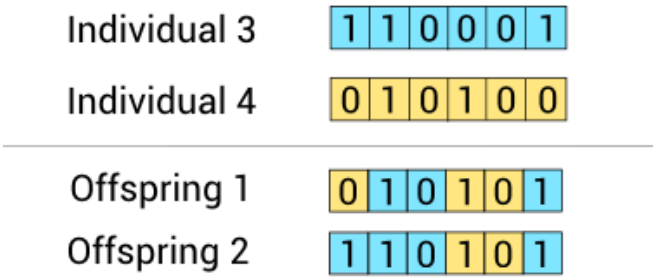
یکی از مهم ترین مراحل،انتخاب individual های مناسب برای ترکیب شدن و رفتن به نسل بعد است.در این جا از دو شیوه به طور همزمان برای انتخاب individual ها استفاده می کنیم.در ابتدا به کمک elitism selection بهترین individual که دارای بیشترین مقدار fitness است،به طور مستقیم به مرحله بعد می رود.روش دیگر، استفاده از roulette wheel است.در این روش طی یک فرآیند تصادفی از بین سه individual باقی مانده،یک مورد به طور تصادفی انتخاب شده و به مرحله بعد می رود.بدیهی است که individual با مقدار fitness بیشتر، شانس بیشتری برای انتخاب شدن دارد.انتخاب هر یک از روش ها به تنهایی،مشکلاتی از قبیل بررسی نشدن فضای جستجو به طور کامل،همگرا شدن به بهینه محلی،از بین رفتن تنوع پاسخ ها و ... باشد.



به طور مثال در شکل بالا،individual چهارم در ابتدا با بیشترین fitness انتخاب می شود و سپس از بین سه مورد باقی مانده،مورد سوم انتخاب می شود.individual دوم با آن که مقدار fitness بیشتری دارد،ولی به دلیل ماهیت تصادفی الگوریتم،انتخاب نشده است.

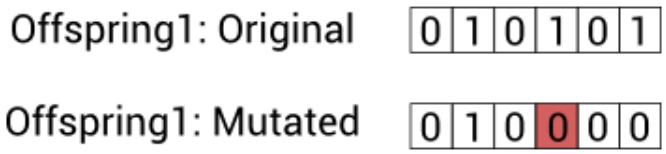
2.4) crossover:

در این مرحله به کمک individual هایی که از مرحله قبل به دست آمد،عملیات crossover را انجام می دهیم. به کمک uniform crossover و احتمال مشخص،مشخص می کنیم که هر ژن از کدام یک از دو والد به ارث برسد.بعضی ویژگی ها از والد اول و بعضی از والد دوم به ارث می رسد.دوبار این عملیات را انجام داده و در نتیجه دو فرزند جدید تولید می شود.



2.5) mutation:

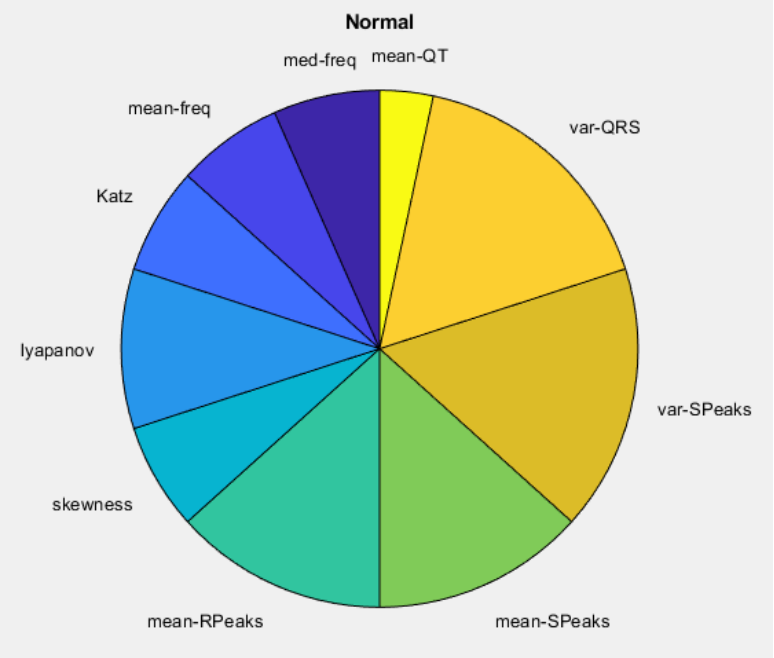
از آنجایی که ممکن است فرزندان تولید شده خیلی شبیه به والدین بوده و این امر باعث کاهش تنوع شود،فلذا از Mutation استفاده می کنیم.در این قسمت،هر ژن از رشته های باینری به دست آمده را با حتمال مشخص عوض می کنیم.این احتمال را برابر با 0.1 در نظر گرفتیم.در نتیجه به طور میانگین،چون individual های ما طول 30 دارند،3 ژن از صفر به یک یا از یک به صفر تبدیل خواهد شد.



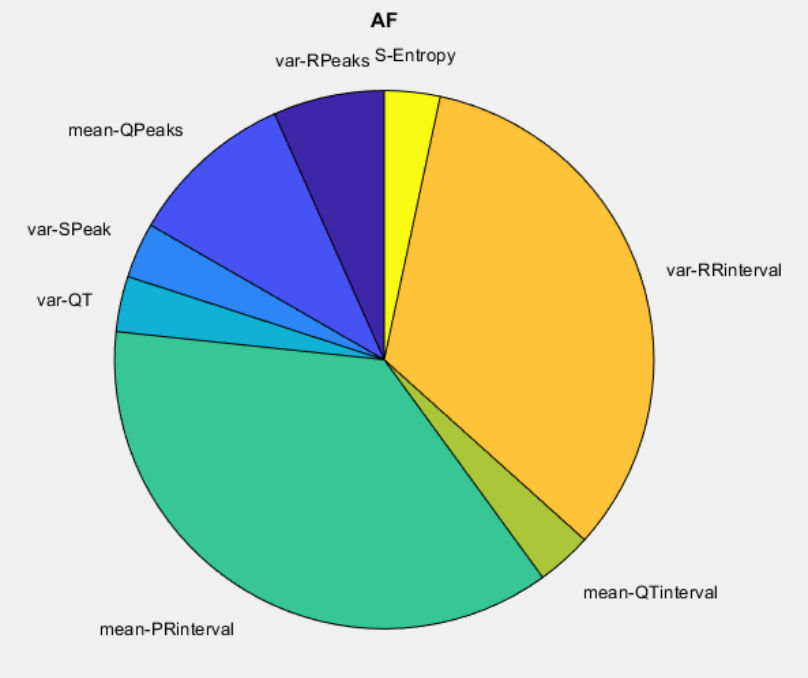
2.6)termination:

برای توقف الگوریتم،راه حل های متنوعی وجود دارد.یکی از راه حل ها می تواند این باشد که اگر خطا برای چند Iteration متفاوت تغییری نکرد و یا مقدار اختلاف خطا برای دو iteration متوالی از مقداری کمتر بود،امکان آن هست که الگوریتم را خاتمه دهیم.یکی دیگر از راه حل ها،سعی و خطا روی iteration های مختلف است به طوری که خروجی بهینه برای ما حاصل شود.ما راه حل دوم را برگزیدیم و پس از سعی و خطا های بسیار،مقداری مناسب که با تعداد جمعیت اولیه نیز متناسب باشد،انتخاب کردیم.

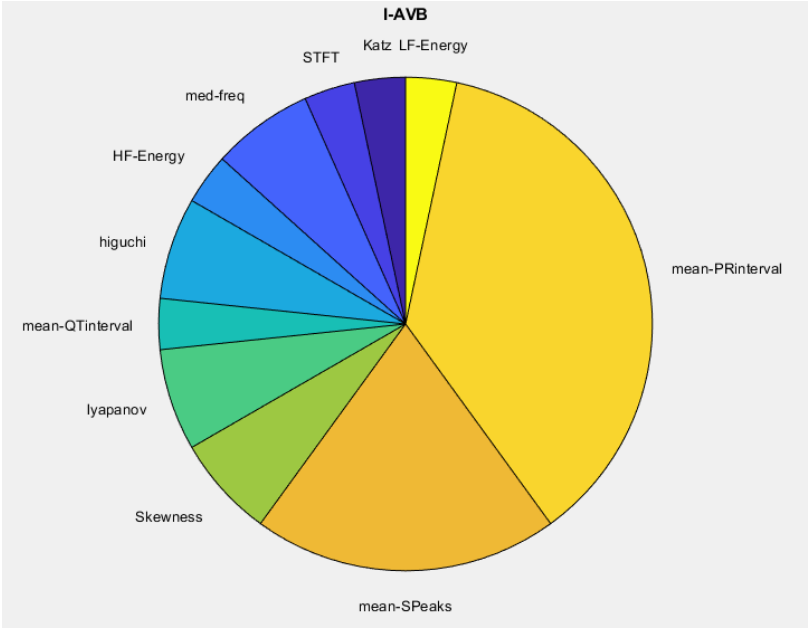
از آن جایی که در فاز اول هم تعداد ویژگی های تاثیر گذار در هر بیماری را 30 در نظر گرفته بودیم،با مقایسه نتایج دیدیم که ویژگی های موثر تا حد بسیار خوبی جدا سازی شده اند که به صورت زیر هستند.



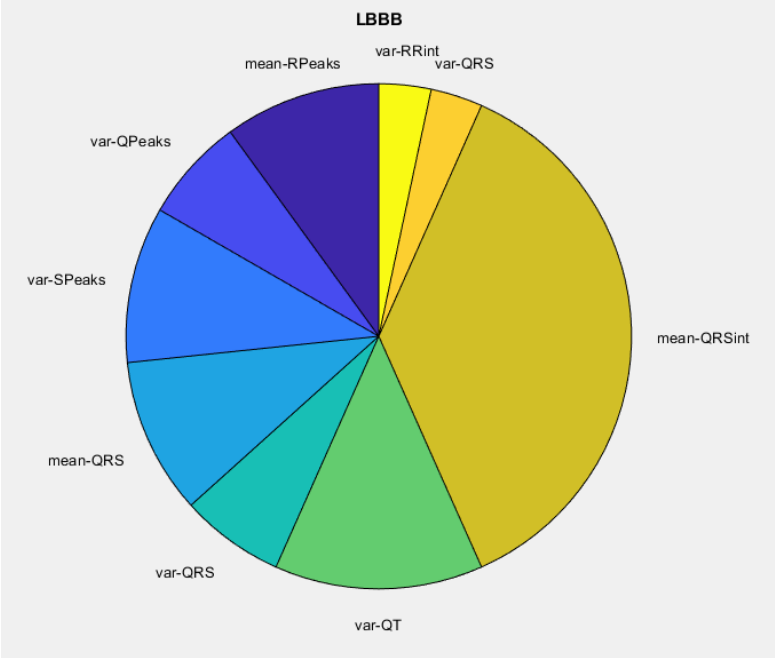
برای حالت نرمال،واریانس پیک S،واریانس QRS،میانگین پیک S ویژگی های مهم و تاثیر گذار برای جدا کردن بیمار ها بودند



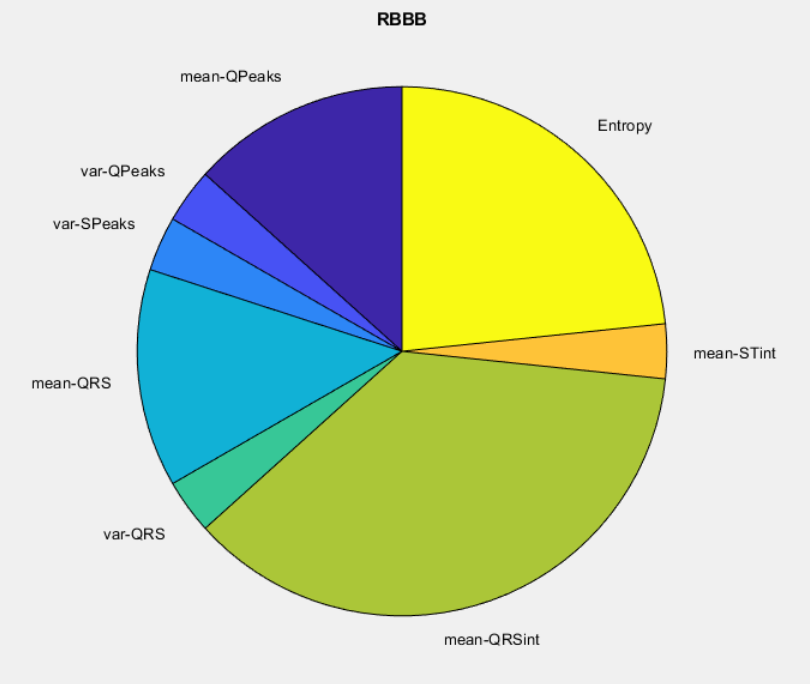
برای AF،مشاهده شد که میانگین بازه RP،نقش بسیار تاثیرگذاری در انتخاب این بیماری دارد به طوری که ویژگی های دوم تا 10 موثر،همگی مربوط به این مورد بودند.همین طور Entropy،ویژگی مهمی بود.



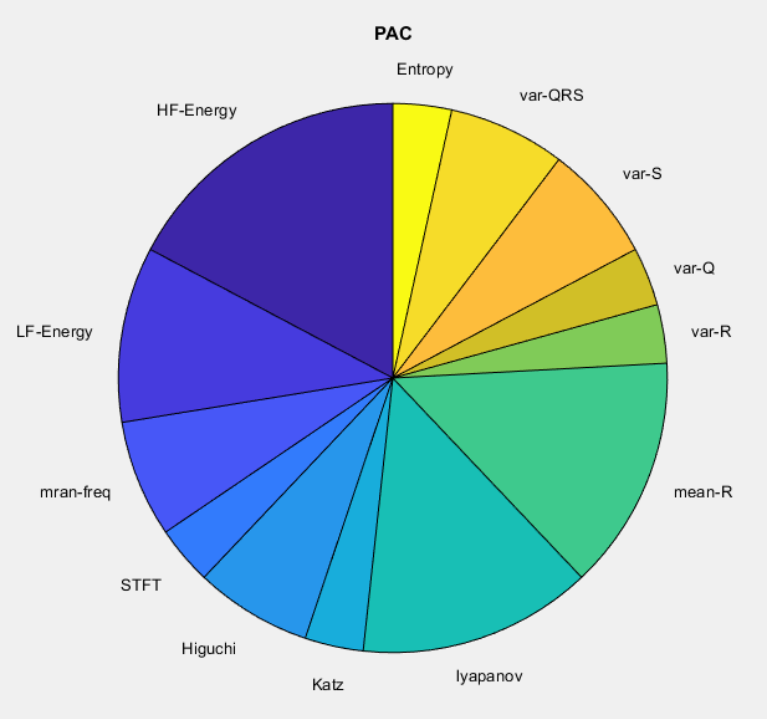
برای I-AVB نیز میانگین مقادیر بازه RP،نقش مهمی در تشخیص این بیماری داشت.



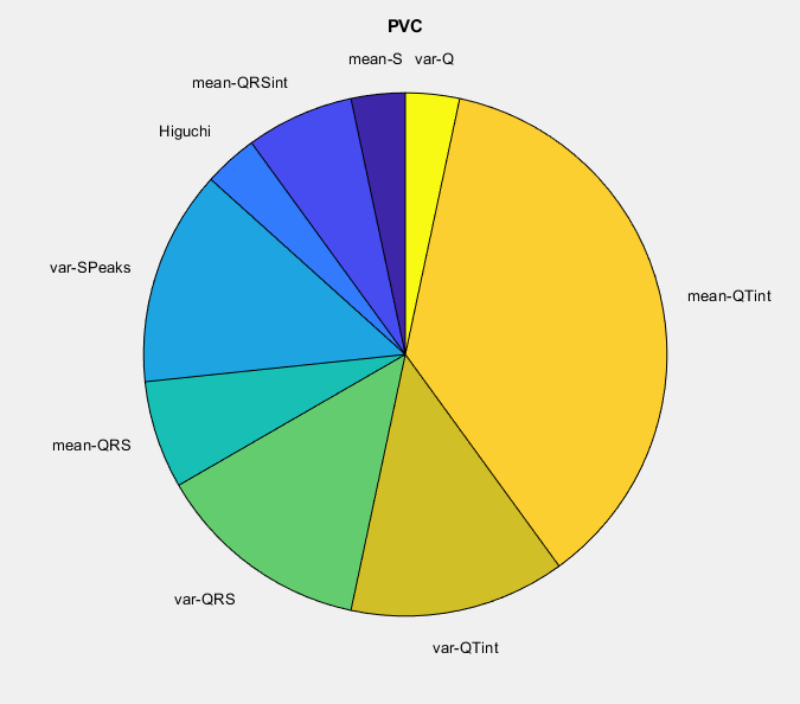
برای LBBB،واریانس بازه QT،واریانس پیک S،واریانس QRS ویژگی های مهم برای تشخیص بودند.



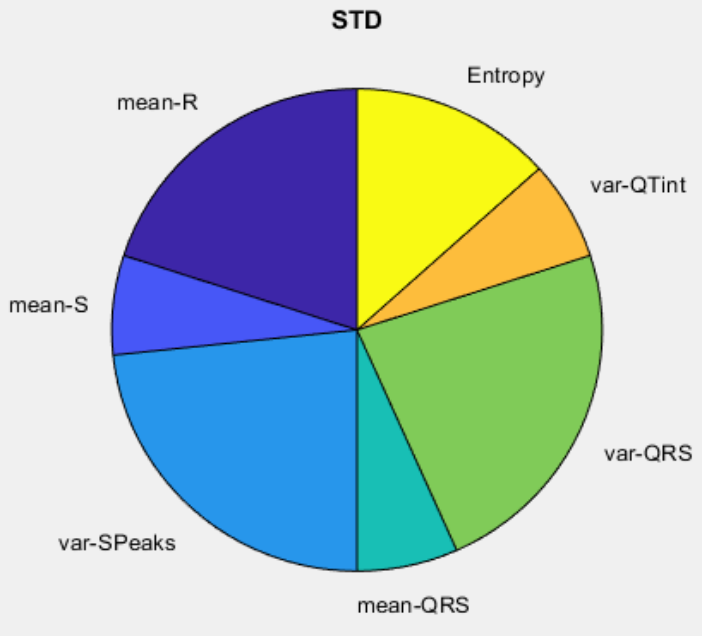
برای RBBB،میانگین مقادیر بازه QTو به خصوص میانگین بازه های زمانی QT بسیار موثر بودند.



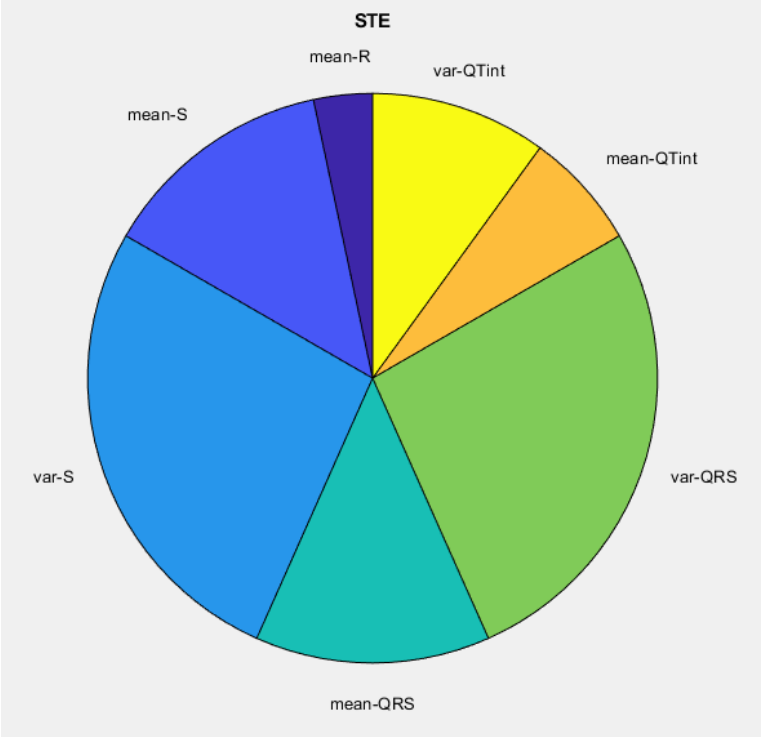
برای PAC،ویژگی های Petrosian،انرژی باندهای فرکانسی مختلف و Katz از ویژگی های مهم بودند.



برای PVC،واریانس بازه QT،واریانس پیک های Q و واریانس پیک S از ویژگی های مهم بودند.



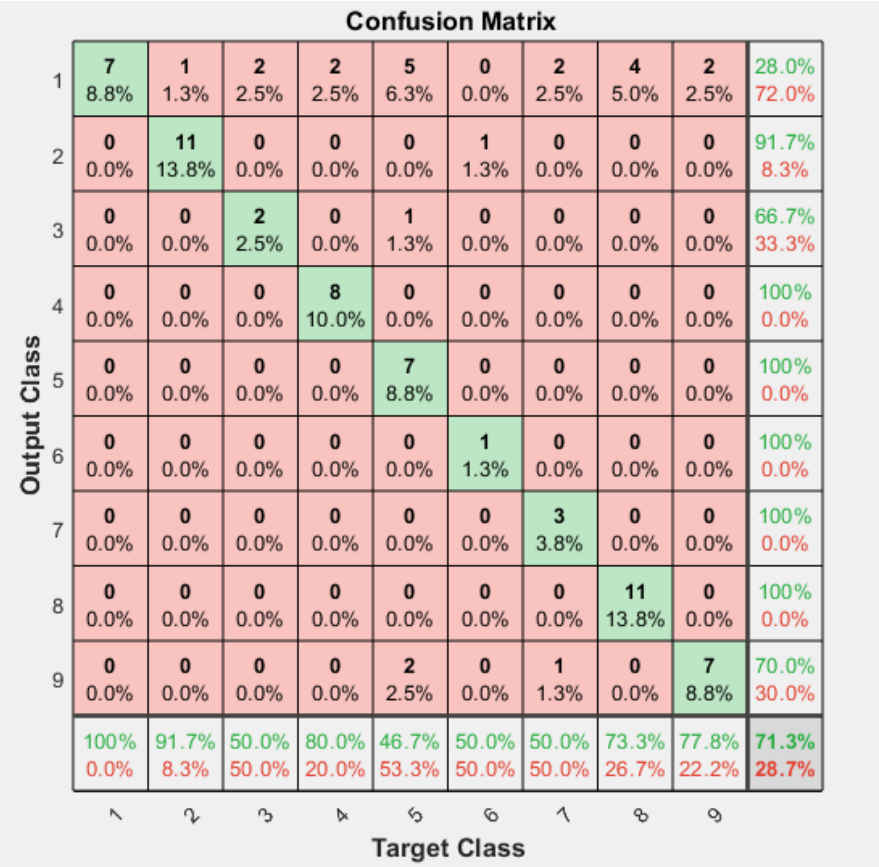
برای STD،واریانس پیک های S و واریانس QRS و Entropy تاثیر گذار بودند.



برای STE،واریانس QRS ، واریانس پیک های S و میانگین بازه QT مهم بودند.

بخش سوم:اعتبار سنجی داده ها

پس از مشخص شدن ویژگی های موثر،به کمک پیاده سازی های فاز دوم،شبکه های RBF را آموزش می دهیم و خروجی برای داده های تست به صورت زیر است.



در فاز قبل،به درصد صحت 82% بر روی داده ها رسیدیم.در این قسمت به کمک ویژگی هایی که به کمک الگوریتم ژنتیک به دست آوردیم،درصد صحت تقریبا برابر با 72% است.

از مزایای الگوریتم های ژنتیک می توان به سهولت پیاده سازی آن ها و عدم نیاز به دانش خاص درباره دیتای اولیه و منعطف پذیر بودن جهت پیاده سازی ایده های متنوع اشاره کرد.از معایب آن می توان گفت که چون ماهیت تصادفی دارند،قطعیتی بر همگرایی آن ها به خروجی مطلوب نیست و این که معمولاً زمان زیادی برای همگرایی و رسیدن به جواب مطلوب نیاز دارند.