تشخیص بیماری آریتمی قلبی با استفاده از روش فیشر و ماشین بردار پشتیبان حداقل مربعات

ایرج نارویی[۱]، بهزاد زمانی[۲]

Diagnosis Of Cardiac Arrhythmia Using Fisher And Ls-Svm

Iraj Naruei, Behzad Zamani

Email: irajnaruei@iauzah.ac.ir

چکیده

بیماریهای قلبی یکی از شایعترین انواع بیماریها است که آمار بسیار بالایی از مرگ و میر را به خود اختصاص میدهد به طوریکه سالانه حدود ۵۰۰۰۰۰ نفر بر اثر این بیماری جان خود را از دست می دهند. آریتمی ها ضربانهای غیر طبیعی هستند، که موجب میشوند قلب خیلی سریع (تاکی کارد) یا خیلی آهسته (برادی کارد) بزند و پمپاژ غیر مؤثر داشته باشد. الکتروکاردیوگرام (ECG) پروسه بدون دردی است که فعالیت الکتریکی قلب را ضبط می کند . تجزیه وتحلیل خودکار ECG برای تشخیص ودرمان بیماران بدحال حیاتی است. ما در این تحقیق از روش فیشر برای استخراج ویژگی های مهم استفاده کردیم و ویژگی های استخراج شده را به عنوان ورودی به طبقه بند Ls-Svm وارد کردیم که برای طبقه بندی دودویی به دقت ۹۷.۱۴٪ و برای طبقه بندی چندکلاسه به دقت ۹۰.۷۱٪ دست یافتیم. این آزمایشات روی مجموعه داده های آریتمی پایگاه UCI انجام شده است.

کلمات کلیدی

بیماری قلبی، الگوریتم فیشر، ماشین بردار پشتیبان، ECG، Ls-Svm

۱. مقدمه

آریتمی ها خیلی شایع بوده و سالانه میلیون ها نفر را در جهان درگیر میکنند.آنها علت اصلی مرگ ناگهانی قلبی در ایالات متحده هستند و سالیانه موجب ۴۰۰۰۰۰ مرگ میشوند. [1] بنابراین آریتمیها میتوانند تهدید کننده حیات باشند، البته در صورتیکه کاهش شدیدی در عملکرد پمپاژی قلب ایجاد نمایند. وقتی عملکرد پمپاژی قلب به مدت بیش از چند ثانیه به شدت کاهش یافت، گردش خون ضرورتا قطع میشود و آسیب به ارگانها (مثل مغز) در عرض چند دقیقه بوجود میآید. در زمینه تشخیص آریتمی کارهای زیاد انجام شده که آقای صمد [۳] و همکارانش در سال ۲۰۱۴ از سه مدل K-nn ، Decision Tree و Naive Bayes برای طبقه بندی استفاده کردند که نرخ طبقه بندی به ترتیب ۶۶.۹۶٪ ، ۵۹.۷۷٪ و ۴۵.۸۵٪ حاصل شد. [2] آقای جدهو [۴] و همکارانش در سال ۲۰۱۰ یک شبکه عصبی مدولار با تعداد لایه های پنهان مختلف از یک تا سه و با درصد آموزش مختلف در پارتیشن مجموعه داده ارائه

---

[۱] دانشجوی کارشناسی ارشد دانشگاه آزاد اسلامی واحد زاهدان

[۲] دکترای هوش مصنوعی و عضو هیات علمی دانشگاه علم و صنعت ایران

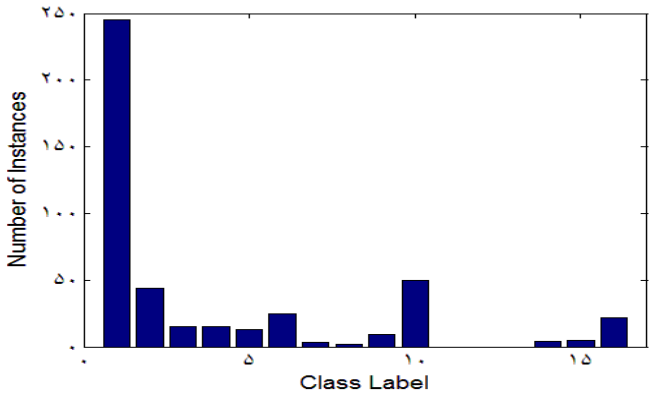
[۳] Samad

[۴] Jadhav

دادند، مقادیر خصیصه های گم شده این مجموعه داده ها با مقادیر نزدیک ترین ستون کلاس مربوطه جایگزین شد. بهترین دقت طبقه بندی با ۲ لایه پنهان برای مجموعه داده پارتیشن ۵ برابر با ۸۲.۲۲٪ شد.[3] آقای جدهو و همکارانش در سال ۲۰۱۰ دوباره یک شبکه عصبی چند لایه با الگوریتم پس انتشار ارائه کردند که به دقت ۸۶.۶۷٪ رسیدند.[4] خانم مالای[۵] و سامانتا در سال ۲۰۱۳ با استفاده از انتخاب ویژگی مبتنی بر همبستگی (CFS) ابعاد ویژگی را کاهش داده و برای طبقه بندی از شبکه عصبی پس انتشار تدریجی و الگوریتم LM استفاده کردند که برای هر طبقه بند به ترتیب به دقت ۸۶.۰۲٪ و ۸۷.۷۱٪ رسیدند.[5] آقای کوهلی[۶] و ورما در سال ۲۰۱۱ از PCA برای استخراج ویژگی و از SVM به عنوان طبقه بند برای ۶ کلاس استفاده کردند که به دقتی بین ۸۲.۵۳٪ تا ۸۵.۷۱٪ دست یافتند. [6]

۲.مجموعه داده

مجموعه داده مورد استفاده در این تحقیق، از آرشیو مجموعه دادههای یادگیری ماشین UCI گرفته شده است.[7] این مجموعه داده شامل ۴۵۲ نمونه و ۲۷۹ ویژگی است که برای سهولت استفاده از آن در تعیین وجود یا عدم وجود بیماری آریتمی و برای شناسایی نوع آریتمی به سه دسته کلی تقسیم شده است. در این مجموعه داده کلاس "۱" اشاره به ECG نرمال، کلاسهای "۲-۱۵" اشاره به کلاسهای مختلف آریتمی و کلاس "۱۶" به بقیه دادههای طبقهبندی نشده اشاره دارد. توزیع نمونه ها در هر کلاس در شکل ۱ نشان داده شده است.



شکل ۱ : تعداد نمونه ها در هر کلاس

۳.پیش پردازش داده ها

در بسیاری از کاربردهای دنیای واقعی کاوش داده ها، حتی با وجود مقدار داده های حجیم و فضای ذخیره سازی مناسب، ممکن است در نمونه های موجود، مقادیری از داده ها از دست رفته(گمشده) باشند. اما مشکل از آنجا آغاز میشود که برای مجموعه داده های بزرگ نمی توان از مقادیر از دست رفته چشم پوشی کرد. یک راه حل برای جایگذینی خودکار مقادیر از دست رفته با مقادیر ثابت عبارت است از:

- جایگزینی تمام مقادیر از دست رفته با یک تک مقدار ثابت سراسری
- جایگزینی یک مقدار از دست رفته با متوسط مشخصه آن.
- جایگزینی یک مقدار از دست رفته با متوسط مشخصه آن برای یک گروه مشخص.

از آنجایکه در دیتاست ما ویژگی هایی با مقادیر مفقودی وجود دارد، و ما نمیخواهیم این اطلاعات را از دست دهیم، لذا مقادیر مفقودی با استفاده از روش میانگین مقادیر هر ویژگی مقدارگذاری می گردد.[8]

---

[۵] Malay
[۶] Kohli

۴.استخراج ویژگی

مسئله عمومی برای انتخاب ویژگی نظارت شده به این شکل است: دیتاست $\{X_i, y_i\}_i^N = 1$ که $X_i \in \mathbb{R}^d$ و $y_i \in \{1,2,...,c\}$ است، هدف ما پیدا کردن یک زیر مجموعه ویژگی با اندازه m که شامل مفیدترین ویژگی هاست. ما از $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d \times n}$ برای نشان دادن ماتریس دادها استفاده می کنیم. X^j نشان دهنده j امین ردیف X است. ۱ یک بردار از همه آنهایی که طول مناسب دارند است. ۰ یک بردار از همه صفرها است. I یک ماتریس واحد با اندازه مناسب است. بدون از دست دادن کلیت، فرض می کنیم که X محور با میانگین صفر شده است، به عنوان مثال، xi = 0 $\sum n_i = 1$ ایده اصلی امتیاز فیشر برای پیدا کردن یک زیر مجموعه از ویژگی هاست به طوری که در فضای داده توسط انتخاب ویژگی ها گرفته شده اند، فاصله بین نقاط داده درکلاس های مختلف تا حد ممکن بزرگ می باشد، در حالی که فاصله بین نقاط داده ها در همان کلاس تا حد امکان کوچک می باشد.

به طور خاص با توجه به m ویژگی انتخاب شده ماتریس داده های ورودی از $X \in \mathbb{R}^{d \times n}$ به $Z \in \mathbb{R}^{m \times n}$ کاهش می یابد، آنگاه امتیاز فیشر به صورت زیر محاسبه می شود.

$$F(Z) = tr\left\{(\tilde{S}_b)(\tilde{S}_t + \gamma I)^{-1}\right\}$$

که γ پارامتر تنظیم مثبت است، $\tilde{S}_b$ ماتریس پراکنده بین کلاسی نامیده می شود، و $\tilde{S}_t$ ماتریس پراکندگی کلی نامیده می شود، که به صورت زیر تعریف شده اند،

$$\tilde{S}_b = \sum_{K=1}^{c} n_k (\tilde{\mu}_k - \tilde{\mu})(\tilde{\mu}_k - \tilde{\mu})^T$$

$$\tilde{S}_t = \sum_{K=1}^{n} (z_i - \tilde{\mu})(z_i - \tilde{\mu})^T$$

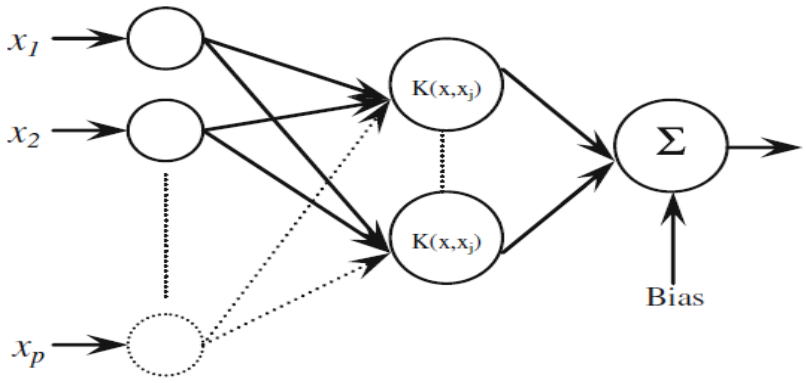
که $\tilde{\mu}_k$ و n_k میانگین بردار و اندازه k مین کلاس نشان داده شده در فضای داده کاهش یافته هستند، به عنوان مثال Z, $\tilde{\mu} = \sum_{K=1}^{c} n_k \tilde{\mu}_k$ میانگین کلی بردار داده های کاهش یافته است. از آنجایکه $\tilde{S}_t$ منحصر به فرد است ما واژه انحراف γI را به نمیه-قطعی مثبت اضافه کردیم. چون $\binom{d}{m}$ کاندید Z خارج از X وجود دارد، مسئله انتخاب ویژگی یک بهینه سازی ترکیبی و بسیار چالش بر انگیز است. برای کاهش مشکلات، به طور گسترده ای استراتژی اکتشافی برای محاسبه امتیاز برای هر ویژگی به طور مستقل بر اساس معیار F مورد استفاده قرار گرفته است. به عبارت دیگر آن تنها $X^j \in \mathbb{R}^{1 \times n}$ در نظر می گیرد. در این موارد فقط $\binom{d}{1} = d$ کاندید وجود دارد. به طور خاص، μ_{jk} و σ_{jk} میانگین و انحراف معیار k امین کلاس مربوط به j امین ویژگی هستند، μ_j و σ_j میانگین و انحرف معیار مجموعه داده مربوط به j امین ویژگی است. بنابراین امتیاز فیشر j امین ویژگی به صورت زیر محاسبه می شود،

$$F(X^j) = \frac{\sum_{k=1}^{c} n_k (\mu_k^j - \mu^j)^2}{(\sigma^j)^2}$$

که $(\sigma^j)^2 = \sum_{k=1}^{c} n_k (\sigma_k^j)^2$ است. پس از محاسبه امتیاز فیشر برای هر ویژگی، m ویژگی با امتیاز بالاتر انتخاب می شود. از آنجا که امتیازهر ویژگی به طور مستقل محاسبه شده، ویژگی های انتخاب شده توسط الگوریتم اکتشافی کمتر از حد مطلوب است. از همه مهمتر، ما قبلا گفتیم، الگوریتم اکتشافی برای انتخاب ویژگی هایی که امتیاز فردی نسبتا پایینی دارند ولی یک امتیاز خیلی بالا زمانی که آنها با یکدیگر ترکیب کامل شده اند دارند شکست می خورد.[9]

۵.ماشین بردار پشتیبان حداقل مربعات (LS-SVM)

سویکنز و وندوال (۱۹۹۹) ماشین بردار پشتیبان با بهینه سازی حداقل مربعات (LS-SVM) را معرفی کردند که فرمول بندی آن محدودیت مشابهی با SVM را بکار می برد. فایده اصلی این روش این است که از نظر محاسباتی بهتر از SVM عمل می کند. در این حالت، آموزش نیاز به حل یک مجموعه توابع خطی بجای مسئله برنامه نویسی دوگانه که حل کلاسیک SVM می باشد، دارد. این روش بطور موثری پیچیدگی الگوریتم را کاهش می دهد. در حالی که در روش SVM از بردارهای پشتیبان برای آموزش و حل مسئله رگرسیونی استفاده می شود، در روش LS-SVM همه داده های آموزشی برای حل مسئله بهینه سازی و تولید نتایج استفاده می شود. در روش LS-SVM داده های آموزشی به فضای هسته ای نگاشت می شوند و برای برقراری توازن بین خطاهای آموزشی و تابع هموار از پارامترهای تنظیم استفاده می کند که برای همه نمونه ها یکسان است و می تواند بعنوان یک پیش فرض در نظر گرفته شود. روش LS-SVM قادر به حل هر دو مسئله طبقه بندی و رگریسون می باشد. [10] معماری این روش بصورت مقابل می باشد:



شکل۲: معماری Ls-Svm

چارچوب حداقل مربعات ماشین بردار پشتیبان به صورت زیر توضیح داده می شود:

مجموعه داده های $\{X_i, y_i\}_i^N = 1$ داده شده است که $X \in \mathbb{R}^p$ به عنوان بردار ورودی و $y_i \in \mathbb{R}$ به عنوان بردار خروجی معرفی می شوند. LS-SVM نیازمند حل مسئله مینیمم سازی زیر است:

$$\min_{w,e,b} j(w,e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2$$

بر اساس محدودیت:

$$y_i = w^T \varphi(x_i) + b + e_i$$

γ پارامتر تنظیم کننده خطا و e میزان خطا را نشان می دهد. حل با استفاده از شکل لاگرانژی از تابع هدف اصلی:

$$L(w,b,e,\alpha) = j(w,e) - \sum_{i=1}^{N} \alpha_i \{w^T \varphi(x_i) + b + e_i - y_i\}$$

α_i ضریب لاگرانژ است. بر اساس شرایط کان-تاکر (KKT):

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i \varphi_i(x_i)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0$$

$$\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i, i = 1,\ldots,N$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow y_k = w^T \varphi(x_i) + b + e_i, i = 1,\ldots,N$$

با شرایط KKT مدل حداقل مربعات ماشین بردار پشتیبان به منظور تابع تخمین به صورت رابطه زیر نتیجه می گردد:

$$y(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i) + b$$

$k(x_i, x_j)$ تابع کرنل نامیده می شود که با تبعیت از شرایط Mercer به عنوان تابعی با ایجاد ضرب داخلی در فضای ویژگی معرفی می شود.

$$k(x_i, x_j) = (\varphi(x_i).\varphi(x_j)), i, j = 1,\ldots,N$$

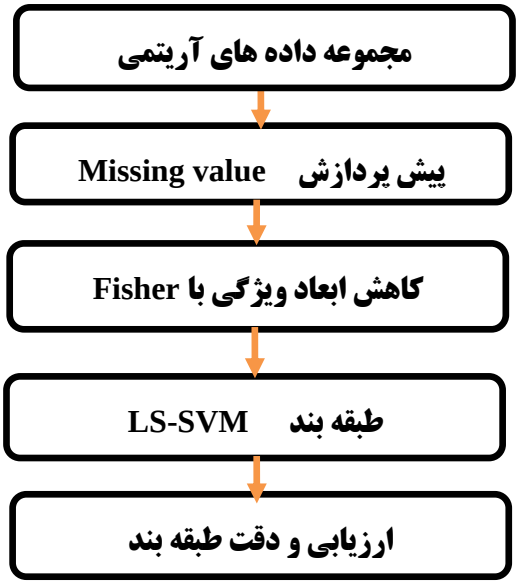
برای حل، تعیین پارامتر تنظیم کننده γ و تعیین σ^2 پارامتر مربوط به تابع کرنل (مربوط به کرنل RBF) نیاز است. انتخاب بهترین تابع کرنل به وسیله سعی و خطا امکان پذیر است. در این تحقیق ما از تابع زیر استفاده کردیم.

$$k(x_i, x_j) = \exp\left(\frac{-\left|x_i - x_j\right|^2}{2\sigma^2}\right)$$

در نهایت طبقه بند LS-SVM با استفاده از حل مجموعه معادلات خطی به صورت زیر عمل می کند:

$$f(x) = sign\left(\sum_{i=1}^{N} y_i \alpha_i k(x, x_i) + b\right)$$

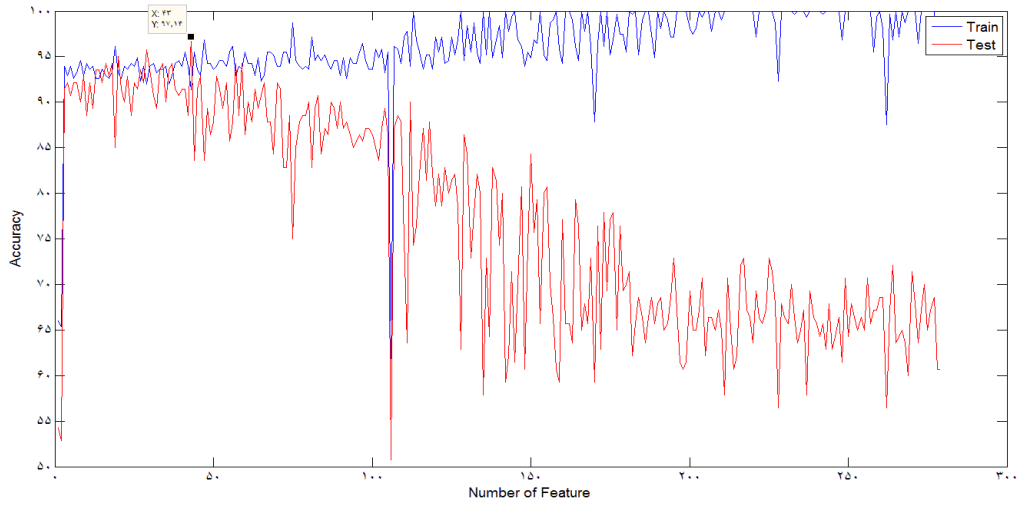
معماری مدل پیشنهادی ما در شکل ۳ ارائه شده است.



شکل ۳: مدل پیشنهادی

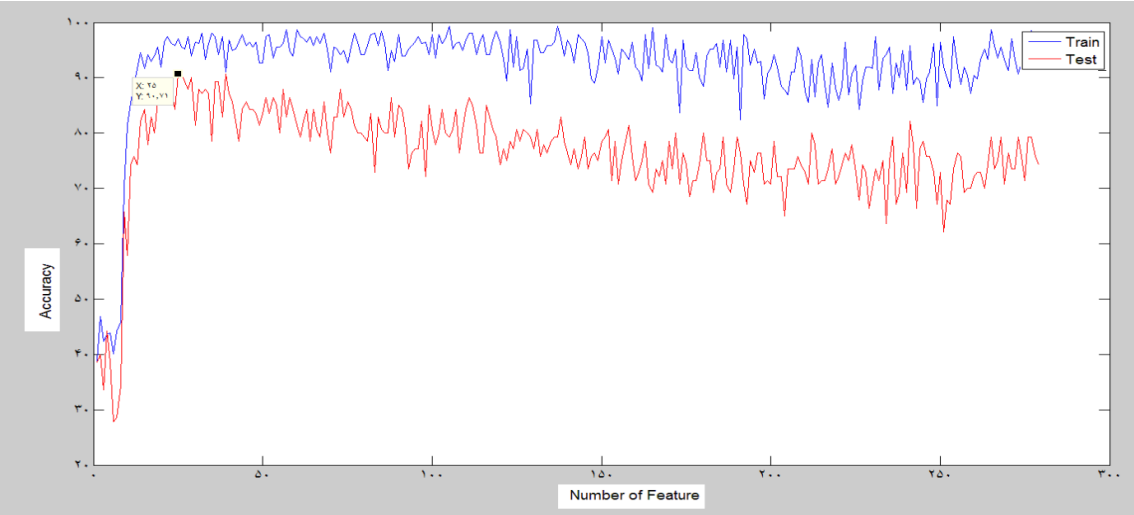
۶.نتایج

ما با استفاده از الگوریتم فیشر که یکی از روش های متاآنالیز است و Toolbox LS-SVM طبقه بندی دوکلاسه و طبقه بندی چند کلاسه داده ها را انجام دادیم، که ۷۰٪ داده ها را برای آموزش و ۳۰ درصد داده ها را برای تست در نظر گرفتیم که برای طبقه بندی دو کلاسه از ۴۳ ویژگی با امتیاز بالاتر بدست آمده از الگوریتم فیشر استفاده کردیم و به دقت ۹۸٪ دست یافتیم. شکل ۴ تاثیر تعداد ویژگی های استخراج شده را روی عملکرد مدل LS_SVM برای طبقه بندی دودویی نشان می دهد.



شکل۴: تاثیر تعداد ویژگی بر دقت الگوریتم کلاس دودویی

برای طبقه بندی چند کلاسه از ۲۶ ویژگی استفاده کردیم و به دقت ۹۷٪ برای داده های آموزش و به دقت ۹۰.۷۱٪ برای داده های تست دست یافتیم. شکل ۵ تاثیر تعداد ویژگی های استخراج شده را روی عملکرد مدل LS_SVM برای طبقه بندی چند کلاسه نشان می دهد.



شکل ۵: تاثیر تعداد ویژگی بر دقت الگوریتم چند کلاسه

در جدول ۱ نتایج ما با روش های دیگر مقایسه شده اند که روش پیشنهادی ما بالاترین دقت را دارد.

جدول ۱ : مقایسه روش ها

ردیف	روش	دقت طبقه بندی دودویی	دقت طبقه بندی چندکلاسه
۱	صمد و همکارانش سال ۲۰۱۴		
	K-nn	۶۶.۹۶	
	Decision Tree	۵۹.۷۷	
	Naïve Bayes	۴۵.۸۵	
۲	جدهو و همکارانش سال ۲۰۱۰		
	شبکه عصبی مدولار	۸۲.۲۲	
۳	جدهو و همکارانش سال ۲۰۱۰	۸۶.۶۷	
	شبکه عصبی چند لایه		
۴	مالای و سامانتا سال ۲۰۱۳		
	شبکه عصبی پس انتشار تدریجی	۸۶.۰۲	
	الگوریتم LM	۸۷.۷۱	
۵	کوهلی و ورما سال ۲۰۱۱		
	SVM-PCA برای ۶ کلاس		۸۲.۵۳ تا ۸۷.۷۱
۶	random forest	۹۰	۷۶
۷	**روش پیشنهادی ما**	**۹۷.۱۴**	**۹۰.۷۱**

۷.نتیجه گیری

این نتایج مبین بهبودی قابل ملاحظه ای نسبت به کارهای اخیر در زمینه تفکیک آریتمی های قلبی است. توجه به این نکته ضروری است که اکثر مقالات یاد شده در بخش منابع بر روی همین پایگاه داده کار کرده اند و بیشتر آنها فقط به تفکیک دو کلاسه پرداخته اند، که با این وجود با نتایج این تحقیق قابل مقایسه نیستند. بکار گیری روش پیشنهادی می تواند کمک شایانی در زمینه تشخیص دقیق و صحیح امراض قلبی به پزشکان نماید.

۸.منابع و مراجع

[1] Sesselberg.HW, et al., 2007,Ventricular arrhythmia storms in post infarction patients with implantable defibrillators for primary prevention indications: A MADIT-II sub study,Heart Rhythm, Elsevier, Vol.4,1395–1402.

[2] Saleha Samad, et al.,2014, Classification of Arrhythmia, International Journal of Electrical Energy, Vol. 2, No. 1.

[3] Shivajirao M. Jadhav, et al., 2010a, ECG Arrhythmia Classification using Modular Neural Network Model, IEEE EMBS Conference on Biomedical Engineering & Sciences.

[4] Shivajirao M. Jadhav, et al., 2010b, Artificial Neural Network Based Cardiac Arrhythmia Classification Using ECG Signal Data, International Conference on Electronics and Information Engineering.

[5] Malay Mitra and R. K. Samanta, 2013, Cardiac Arrhythmia Classification Using Neural Networks with Selected Features, International Conference on Computational Intelligence, 76– 84.

[6] Narendra Kohli and Nishchal K. Verma,2011, Arrhythmia classification using SVM with selected features, International Journal of Engineering, Science and Technology Vol. 3, No. 8, 2011, pp. 122-131

[7] Bache, K. and Lichman M., 2013, UCI machine learning repository,. URL http://archive.ics.uci.edu/ml.

[8] Jose M. Jerez, et al.,2010, Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artificial Intelligence in Medicine vol. 50, pp. 105–115.

[9] Q. Gu, et al.,2011, Generalized fisher score for feature selection, in Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI '11), pp. 266–273.

[10] A. Mellit, et al.,2013, Least squares support vector machine for short-term prediction of meteorological time series, Theor Appl Climatol vol.111, pp. 297–307.