

تحلیل لایه بندی ژنتیکی تیره های اسب عرب ایرانی به روش هوش مصنوعی (یادگیری ماشین)

به‌هکام تیموری^۱، حسین مرادی شهر بابک^{۲*}، محمد مرادی شهر بابک^۳، میر باقر زندی باغچه مریم^۴، علیرضا فتوحیسیاه پیرانی^۵

۱، ۲ و ۳) گروه علوم دامی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران.

۴) گروه علوم دامی، دانشگاه زنجان، زنجان، ایران.

۵) استادیار موسسه بیوشیمی و بیوفیزیک، دانشگاه تهران، تهران، ایران.

شماره صفحات

۳۶-۲۵

نویسنده مسئول: hmoradis@ut.ac.ir

تاریخ دریافت: ۱۴۰۴/۰۴/۲۵

تاریخ پذیرش: ۱۴۰۴/۰۷/۳۰

چکیده

اسب نژاد عرب یکی از کهن‌ترین و اصیل‌ترین نژادهای اسب در دنیا است که به دلیل ویژگی‌هایی منحصر به فرد، جایگاه ویژه‌ای در اصلاح و توسعه نژاد های دیگر دارد. اسب عرب ایرانی (اصیل) یکی از مهم‌ترین منابع ژنتیکی دام در کشور است که شناسایی ژنتیکی تیره‌های آن برای حفظ تنوع و اصلاح نژاد حیاتی می‌باشد و زمینه ارتقای برنامه‌های اصلاح نژاد، توسعه و حفظ این نژاد را فراهم می‌کند. به منظور مطالعه ساختار ژنتیکی تیره های مختلف اسب عرب ایرانی، داده‌های ژنومی ۱۰۹ راس اسب عرب ایرانی از شش تیره (خرسان: ۵۳، حمدانی: ۱۹، عیبیان: ۱۳، صغلاوی: ۱۲، کحیلان: ۹، شویمان: ۳) و ۳۸،۵۷۲ نشانگر SNP پس از انجام مراحل کنترل کیفیت با نرم افزار PLINK و شاخص های $(MAF > 0.03)$ ، $(MIND > 0.03)$ ، $(p > 10^{-7})$ تجزیه و تحلیل قرار گرفت. پس از کاهش ابعاد با روش تحلیل مؤلفه‌های اصلی (PCA)، الگوریتم‌های یادگیری ماشین نظارت‌شده که شامل جنگل تصادفی (Random Forest)، ماشین بردار پشتیبان (SVM) و پرسپترون چندلایه (MLP)، استفاده شدند. آزمون فریدمن نشان داد که مقدار p -value برابر با ۰.۰۰۰۱۹ است و تفاوت معناداری بین مدل‌ها وجود دارد و انتخاب مدل را تحت تاثیر قرار می‌دهد. مدل SVM با میانگین دقت (0.83) نسبت به MLP (0.73) و RF (0.75) عملکرد بهتری نشان داد و اختلاف عملکرد بین مدل‌ها مشاهده شد. این یافته‌ها کاربرد مدل‌های هوش مصنوعی یادگیری ماشین به‌ویژه ماشین بردار پشتیبان (SVM) را در طبقه‌بندی مؤثر تیره‌های اسب عرب با داده‌های ژنومی تایید می‌کند که می‌تواند به بهبود برنامه‌های اصلاح نژاد کمک شایانی نماید.

کلمات کلیدی: اسب عرب ایرانی، یادگیری ماشین، جنگل تصادفی، ماشین بردار پشتیبان و پرسپترون چندلایه.

مقدمه

اسب عرب به عنوان یکی از قدیمی‌ترین و اصیل‌ترین نژادهای اسب در جهان، از دیرباز مورد توجه بوده و نقش مهمی در تاریخ و فرهنگ بشریت ایفا کرده است. این نژاد با ویژگی‌های منحصربه‌فردی همچون زیبایی، هوش، استقامت و سرعت، در سراسر جهان پراکنده شده و به عنوان پایه بسیاری از نژادهای دیگر مورد استفاده قرار گرفته است (Hill et al., 2012). اگرچه این اسب در تاریخ به نام اسب عرب شهرت یافته، اما ریشه و پیشینه آن به ایران و قبایل مختلف بادیه نشین حاشیه فالت ایران و شبه جزیره عربستان بازمی‌گردد (Bashardoost, 2020). اسب عرب به پنج زیرخانواده اصلی یعنی کُحیلان، هدبان، حمدانی، عُبَیان و صَعْلَوی طبقه بندی می‌شود. برخلاف دام‌های دیگر که جهت ثبت شجره و خط خونی نام پدر اهمیت دارد در اسب عرب نام مادین در نام گذاری تیره‌ها مورد استفاده قرار می‌گیرد که در مطالعه ای فرضیه منشأ تبار مادری اسب عرب توسط داده‌های ژنتیکی تأیید نشد و اعلام کرد که تاریخچه این نژاد را باید حاصل فرایند بلندمدت مهاجرت، تبادل و انتخاب دانست (Głazewska, 2010). از بین تمام تیره‌های اسب نژاد عرب، تیره کوهیلان ویژگی‌های «مردانه» و ساختاری قوی دارد، تیره صقلای ویژگی‌های «زنانه» با شکل ظریف تر و تیره مونیقی را به دلیل سرعت بالای آن به عنوان یک نوع «سریع» توصیف می‌کرد. این طبقه بندی‌ها عموماً بر مبنای ویژگی‌های ظاهری همچون فرم سر و گردن و حالت دم صورت می‌گیرد. در کل ۲۰ تیره شناخته شده از اسب‌های عربی وجود دارد که تنها دوازده تیره از آن‌ها به صورت خالص حفظ و پرورش داده شده‌اند. در ایران، اسب عرب به عنوان «سب/اصیل/ایرانی» شناسایی شده است (Khaleili, 2008) (Forbis, 1986). این تیره‌ها، علاوه بر تفاوت‌های ظاهری، ممکن است دارای ساختار ژنتیکی متمایزی باشند که مطالعه ی دقیق آنها برای حفظ خلوص نژادی و تنوع ژنتیکی ضروری است. اهمیت اسب عرب نه تنها به دلیل پیشینه تاریخی آن، بلکه به خاطر نقش آن در ژنتیک و اصلاح نژاد دیگر نژادها است. در دهه‌های اخیر، با پیشرفت‌های چشمگیر در حوزه بیوانفورماتیک و ژنومیک، ابزارهای قدرتمندی برای تحلیل داده‌های ژنتیکی حجیم، به ویژه نشانگرهای تک نوکلئوتیدی SNP، توسعه یافته‌اند. این نشانگرها به دلیل فراوانی در ژنوم، به ابزاری کلیدی برای بررسی تنوع ژنتیکی، ساختار جمعیتی و شناسایی مناطق ژنومی مرتبط با صفات مهم در گونه‌های مختلف، از جمله اسب، تبدیل شده‌اند (Ghasemi et al., 2021) (Hill et al., 2021). اگرچه نشانگرهای snp فرصت‌های بی نظیری برای درک عمیق‌تر از ژنتیک فراهم آورده، اما همزمان چالش‌هایی را نیز در زمینه پردازش، تحلیل، و استخراج اطلاعات کاربردی از این داده‌های پیچیده ایجاد کرده است. روش‌های آماری کلاسیک گاهی در شناسایی الگوهای غیرخطی و تعاملات پیچیده ژنومی با محدودیت‌هایی مواجه می‌شوند. در این میان، یادگیری ماشین (Learning Machine) به عنوان یکی از شاخه‌های هوش مصنوعی، با توانایی چشمگیر خود در شناسایی الگوها، روابط پیچیده، و پیش‌بینی از مجموعه داده‌های بزرگ، نقش مهمی در تحلیل داده‌های ژنومی پیدا کرده است (Hastie et al., 2009). الگوریتم‌های یادگیری ماشین، با توانایی بالا در شناسایی الگوهای پنهان و روابط غیرخطی در داده‌ها کاربردهای گسترده‌ای، در زمینه‌های مختلف ژنتیک از جمله،

شناسایی و تفکیک نژادهای دامی، طبقه بندی افراد بر اساس ژنوم، پیشبینی صفات، و مطالعه ساختار جمعیت حیوانات پیدا کرده‌اند (Valdes *et al.*, 2023; Analytics Steps, n.d.). سوابق پژوهش های انجام شده نشان میدهد که روش های یادگیری ماشین، مخصوصاً الگوریتم هایی مانند جنگل تصادفی (Random Forest (RF)، ماشین بردار پشتیبان Support Vector Machine (SVM) و پرسپترون چندلایه (Multi Layer- Perceptron (MLP)، توانسته اند ساختار ژنتیکی و تفکیک جمعیت های مشابه را با دقت بالا انجام دهند (Cervantes *et al.*, 2020; Valdes *et al.*, 2023). به عنوان مثال، جنگل تصادفی به دلیل توانایی در مدیریت داده های پرت و متغیرهای زیاد، و همچنین ارائه معیارهای اهمیت متغیرها، در مطالعات ژنتیکی بسیار محبوب شده است. SVM با یافتن ابرصفحه ای بهینه برای جداسازی کلاس ها، در داده های با ابعاد بالا عملکرد قابل قبولی دارد و MLP نیز به عنوان یک شبکه عصبی پایه، توانایی مدلسازی روابط غیرخطی پیچیده را دارد. با این حال، با وجود پتانسیل بالای این روش ها، ژنتیک اسب های عرب ایران کمتر به طور هدفمند و مبتنی بر داده های کامل SNP با این روشها تحلیل شده است. بسیاری از مطالعات قبلی بر روی اسب عرب ایرانی، به دلیل محدودیت های تکنولوژیکی یا هدف پژوهش، از نشانگرهای میکروستلایت، یا از روش های کلاسیک تکنیک لایه بندی جمعیت استفاده کرده اند که ممکن است قادر به کشف تمام پیچیدگی های ساختار ژنتیکی این نژاد نباشند. لذا، یک تحلیل جامع با استفاده از داده های SNP و روش های پیشرفته یادگیری ماشین، ضروری به نظر میرسد. هدف از این پژوهش، بررسی کارایی روش های یادگیری ماشین نظارت شده در مطالعه ساختار جمعیت های اسب عرب ایرانی با استفاده از داده های ژنومی مبتنی بر نشانگرهای SNP است. در این مطالعه، از الگوریتم های جنگل تصادفی (Random Forest)، ماشین بردار پشتیبان (SVM) و پرسپترون چندلایه (MLP) به منظور مطالعه ساختار جمعیت اسب عرب ایران (اصیل) و تفکیک تیره ها با روش های مبتنی بر تکنیک لایه بندی تیره ها استفاده شده و نتایج حاصل با یکدیگر مقایسه شده اند. هدف این مطالعه، ارزیابی الگوریتم های یادگیری ماشین (Random Forest، SVM و MLP) در شناسایی ساختار جمعیت و تفکیک تیره های اسب عرب ایرانی با استفاده از داده های ژنومی مبتنی بر SNP است.

مواد و روش ها

در این پژوهش، در مجموع، اطلاعات تعیین ژنوتایپ تعداد ۱۰۹ راس اسب نژاد عرب ایران (اصیل) از شش تیره (خرسان: ۵۳، حمدانی: ۱۹، عیبان: ۱۳، صغلاوی: ۱۲، گحیلان: ۹، شویمان: ۳) از داخل کشور، برای این مطالعه اخذ گردید. نمونه های مو اسب ایرانی از استان های مختلف کشور جمع اوری شده است. تمامی نمونه های مورد مطالعه در این تحقیق مراحل آزمایش انساب را گذرانده و از لحاظ خلوص نژادی مورد تایید هستند و در کتاب انساب نژادی مربوطه قرار دارند. همه نمونه ها با استفاده از تراشه های ژنومی Illumina 70k SNP equine bead chip و با بکارگیری دستورالعمل استاندارد شرکت Illumina تعیین ژنوتایپ شد و براساس تجزیه و تحلیل ۳۸,۵۷۲ نشانگر ژنتیکی SNP، مورد استفاده قرار گرفت. کیفیت اولیه داده ها با نرم افزار PLINK نسخه ۱.۹ ارزیابی شد که شاخص های مربوط به کنترل کیفیت شامل حذف SNP های دارای نرخ ژنوتایپینگ پایین

(GENO < ۰.۰۳) حذف نشانگرهای دارای فراوانی آلل حداقل کمتر از ۳ درصد (MAF > ۰.۰۳) حذف SNPهای خارج از تعادل هاردی-واینبرگ ($p > 10^{-7}$) حذف نمونه‌هایی با نرخ تعیین ژنوتایپ کمتر از ۳ درصد (MIND < ۰.۰۳) را شامل می‌شود. همچنین شاخص Fst، به عنوان معیاری برای سنجش فاصله ژنتیکی بین جمعیت‌ها، محاسبه گردید (Purcell et al., 2007). برای کاهش ابعاد بالای داده‌های SNP و شناسایی مهم‌ترین ویژگی‌های ژنتیکی، از تحلیل مؤلفه‌های اصلی Principal Component Analysis (PCA) با بسته نرم‌افزاری R و تابع prcomp استفاده شد. هدف از این روش با تبدیل داده‌ها به چند مؤلفه اصلی، تجسم ساختار ژنتیکی جمعیت و تمایز تیره‌های مختلف اسب عرب ایرانی در فضای دوبعدی (PC1/PC2) است (شکل ۱).

روش های طبقه‌بندی با استفاده از الگوریتم‌های یادگیری ماشین

به منظور طبقه‌بندی و تفکیک تیره های اسب عرب ایران (اصیل)، از سه الگوریتم یادگیری ماشین نظارت‌شده شامل جنگل تصادفی (Random Forest)، ماشین بردار پشتیبان (SVM) و پرسپترون چندلایه (MLP) در محیط Python استفاده شد. داده‌ها پس از کاهش ابعاد با تحلیل مؤلفه‌های اصلی (PCA) و شناسایی ویژگی‌های اصلی، به دو بخش آموزش (۷۰٪) و آزمون (۳۰٪) تقسیم شدند. مدل‌ها توسط تابع run_classifier آموزش داده شد و پیش‌بینی بر روی داده‌های آزمون انجام شد. به منظور نمایش دادن نتایج مدل‌ها، از ماتریس درهم‌ریختگی (Confusion Matrix) توسط تابع make_confusion_heatmap بهره گرفته شد و سپس نتایج به صورت جمع بندی به صورت (جدول ۱) ارائه شد. فرایند ارزیابی برای هر مدل ۱۰ بار تکرار شد و میانگین پیش‌بینی صحیح نتایج به عنوان شاخص نهایی دقت مدل‌ها گزارش گردید (جدول ۲) و سپس برای مقایسه آماری عملکرد مدل‌ها، از آزمون فریدمن استفاده شد که در صورت معنی‌داری ($p < 0.05$) مشخص کننده تفاوت معنی دار بین مدل‌ها است.

جنگل تصادفی

برای مدل‌سازی و طبقه‌بندی داده‌های ژنتیکی در این پژوهش، الگوریتم جنگل تصادفی (Random Forest) به عنوان یکی از روش‌های پرکاربرد یادگیری ماشین مورد استفاده قرار گرفت (Liaw & Wiener, 2002; Ogutu et al., 2011). این الگوریتم با ساخت مجموعه‌ای از درختان تصمیم‌گیری و تجمیع نتایج آن‌ها، قادر است الگوهای پیچیده بین ویژگی‌های ورودی و برچسب‌های خروجی را بیابد (Breiman, 2001). در این مطالعه، مدل Random Forest با ۵۰۰ درخت و تنظیم پارامترهای اولیه، روی بخش آموزش داده، آموزش داده شد و پیش‌بینی کلاس‌بندی روی داده‌های آزمون صورت گرفت. برای تیره‌های اسب عرب به صورت زوجی، مدل مذکور اجرا و عملکرد آن بر اساس میانگین نتایج ده بار تکرار، با استفاده از ماتریس درهم‌ریختگی ارزیابی و مقایسه شد. (Pedregosa et al., 2011).

ماشین بردار پشتیبان

همچنین جهت بررسی عملکرد مدل‌ها روی داده‌های با ابعاد بالا، از الگوریتم ماشین بردار پشتیبان (Support Vector Machine, SVM) با در نظر گرفتن کرنل خطی استفاده شد (Cortes & Vapnik, 1995; Bridges *et al.*, 2011). الگوریتم SVM با هدف یافتن مرزی بهینه جهت تفکیک کلاس‌ها، خصوصاً با توجه به داده‌های ژنتیکی که اغلب دارای هم‌پوشانی هستند، توانایی مناسبی را ارائه می‌دهد (Schölkopf & Smola, 2002). مشابه با روش قبلی، پس از استانداردسازی و کاهش ابعاد داده‌ها با PCA، مدل SVM آموزش و روی مجموعه آزمون، ارزیابی شد. تمامی مراحل برای هر یک از زوج‌های تیره در ده بار تکرار مستقل انجام شد (Pedregosa *et al.*, 2011).

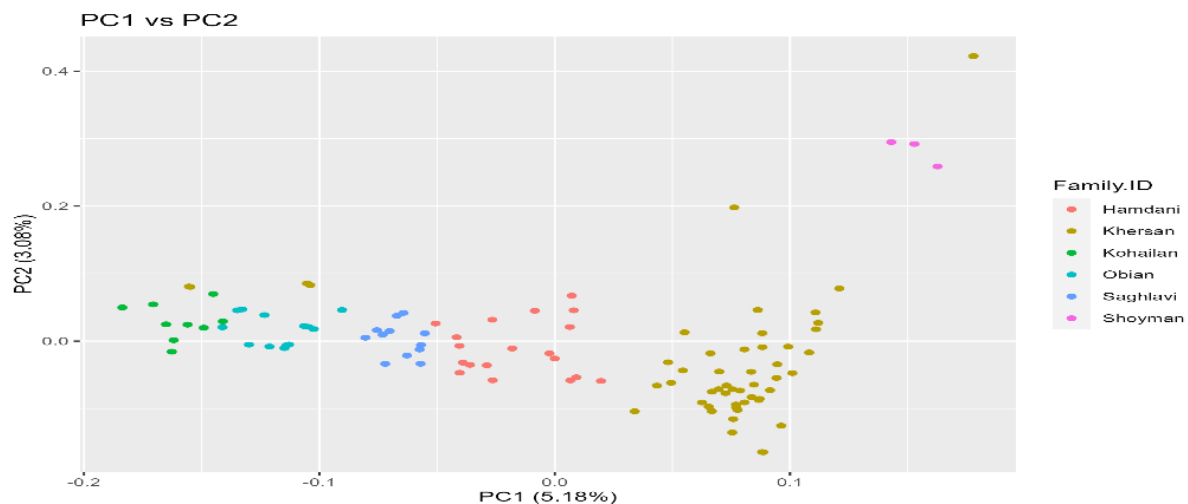
پرسپترون چندلایه

به منظور بهره‌گیری از مدل‌های شبکه‌های عصبی مصنوعی و سنجش توانایی آن‌ها در طبقه‌بندی، از شبکه‌ی عصبی پرسپترون چندلایه (Multi-layer Perceptron, MLP) استفاده شد (Leshno *et al.*, 1993; Motsinger-Reif & Ritchie, 2008). ساختار شبکه شامل یک لایه پنهان با ۱۰ نورون بود. این مدل پس از استانداردسازی و کاهش ابعاد داده‌ها از طریق PCA، مشابه روند سایر مدل‌ها، آموزش دید و ارزیابی شد. تمامی مدل‌ها در چارچوب مشخص و بر بستر یکسان داده‌ای و فرآیند تکرارپذیر مورد مقایسه قرار گرفتند تا از نظر دقت، صحت، و قابلیت تعمیم، سنجش شفافیت از عملکرد آن‌ها به دست آید. همچنین به منظور تحلیل آماری تفاوت مدل‌ها از آزمون فریدمن و در صورت معنی دار بودن تفاوت از post-hoc جهت مشخص کردن تفاوت بین روش آماری استفاده شد. این فرآیند امکان شناسایی تفاوت‌های معنادار میان مدل‌ها را از منظر تشخیص تیره‌های مختلف اسب عرب ایرانی فراهم می‌سازد (Liu *et al.*, 2022).

نتایج

پس از اجرا کنترل کیفی نتایج ژنوتایپ‌ها، داده‌ها بدون هیچ حذف نمونه یا نشانگر باقی ماندند و تعداد کل جفت snpها معادل ۳۸,۵۷۲ مورد استفاده قرار گرفت و فایل‌های خروجی با فرمت‌های استاندارد برای مراحل بعدی ذخیره گردید. در نهایت چندشکلی‌های تک نوکلئوتیدی که از تمام مراحل کنترل کیفیت عبور کردند برای آنالیزهای بعدی نگه داشته شدند. جهت برآورد فاصله ژنتیکی بین تیره‌ها از آماره F_{st} بر روی داده‌های ژنومی اسب‌های عرب ایرانی استفاده شد و نتایج نشان داد که میانگین شاخص F_{st} در بین شش خوشه جمعیتی برابر با ۰.۰۳۰۴ و مقدار وزن دار آن ۰.۰۳۳۸ است. در مطالعه‌ای که بر روی اسب‌های عربی خاورمیانه و غربی انجام شده بود مشخص شد که تمایز میان جمعیت‌های عربی خاورمیانه‌ای بسیار پایین است ($F_{ST} < 0.02$) در حالی که تفاوت قابل توجهی بین عرب‌های خاورمیانه و جمعیت‌های اسب عرب پرورش یافته در غرب مانند اروپا و آمریکا وجود دارد ($F_{ST} < 0.1$) که بر اساس استانداردهای ژنتیکی جمعیت، این مقادیر نشان‌دهنده تمایز ژنتیکی کم تا متوسط بین زیرجمعیت‌های مورد مطالعه است (Khanshour, 2013). نتایج نشان می‌دهد که زیرجمعیت‌های اسب عرب ایرانی تا حدی

از نظر ژنتیکی متمایز هستند که در گزارش مطالعه (Sadeghi *et al.*, 2019) اشاره شد Fst بین زیر گروه های اسب عرب معنا دار بوده است، اما احتمالاً در گذشته جریان ژنی بین آنها وجود داشته است. در این مطالعه از اطلاعات جمعیت اسب عرب ایرانی متشکل از ۶ تیره جهت آنالیز مولفه های اصلی PCA مورد استفاده قرار گرفت. نتایج آنالیز PCA (شکل ۱) نشان داد که جمعیت مورد مطالعه بر اساس دو مؤلفه اصلی PC1 و PC2 امکان جداسازی نسبی در تیره ها را فراهم میکند. نتایج pca به ترتیب ۵.۱۸٪ و ۳.۸۶٪ از کل واریانس ژنتیکی را مشخص کرده و نمودار پراکنش PC1 در مقابل PC2 تمایز ژنتیکی میان

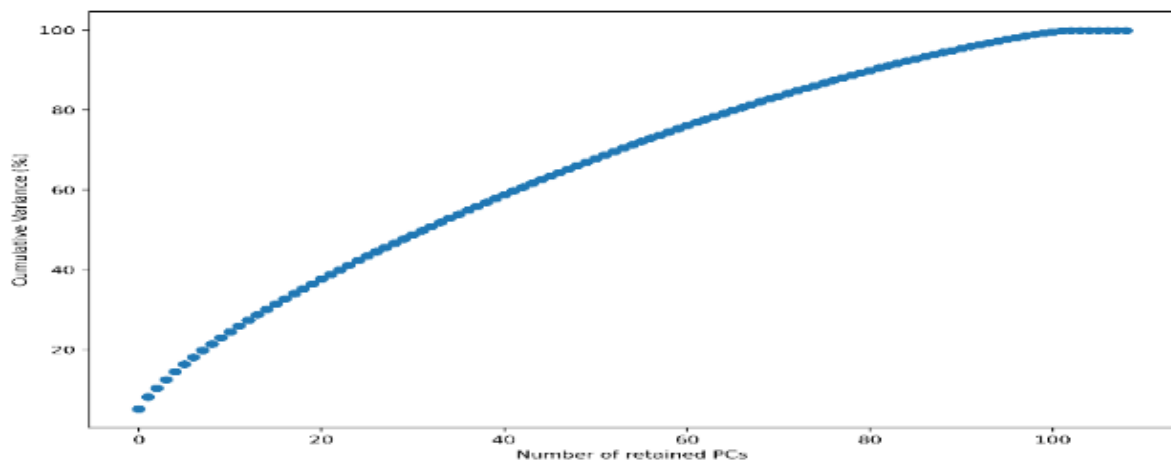


شکل ۱. نمودار پراکنش تیره های مورد بررسی به کمک آزمون تجزیه به مولفه های اصلی PCA

Figure 1. Scatter plot of the studied families using the principal component analysis test (PCA)

برخی تیره ها مانند کُحیلان، عَبیان و صُغلاوی با تیره خرسان را نشان داد، در حالی که نزدیکی ژنتیکی بیشتری بین تیره های صُغلاوی، حمدانی و بین کوهیلان، عَبیان مشاهده میگردد. همچنین، مطابق با واریانس تجمعی، ۶۰ مؤلفه نخست بیش از ۸۰ درصد تغییرپذیری کل داده ها را پوشش می دهند (شکل ۱). و این موضوع بیانگر اثربخشی بالای روش PCA در فشرده سازی داده های ژنومی پیچیده است. در مطالعه دیگری بر اساس pca اسب های نژاد عرب یک خوشه ژنتیکی بزرگ و گسترده تشکیل دادند که از سایر نژادها متمایز بود ولی دامنه تنوع داخلی بالایی داشت و نمونه هایی از نژادهایی مانند کرد و دره شوری ایران را هم در بر میگيرد، همچنین تیره عرب مصری در نمودار PCA به صورت مجزا از اسب های نژاد عرب ایرانی (اصیل)، لهستانی، بحرین، عربستان سعودی و تونس قرار گرفتند (Sadeghi *et al.*, 2019). در مطالعه انجام شده (Badbarin *et al.*, 2022) که با نشانگر های میکروستلایت آنالیز مولفه های اصلی (PCA) و دندروگرام ها نشان داد هیچ گروه بندی ژنتیکی منطبق با تیره ها مشاهده نشد و تمامی تیره ها در یک خوشه بزرگ قرار می گیرند. یعنی تفاوت ژنتیکی بین تیره ها کم و بیشترین تنوع درون جمعیت است. در مطالعه دیگری با استفاده از تحلیل های چندمتغیره، مانند PCA، نشان داد تمامی جمعیت های عرب خاورمیانه (سوری، ایرانی، سعودی) در یک خوشه قرار می گیرند، در حالی که جمعیت های غربی آمریکایی، شایگا و لهستانی خوشه های

جداگانه را تشکیل می‌دهند (Khanshour, 2013). این رویکرد استفاده از pca علاوه بر کاهش پیچیدگی محاسباتی، به آشکارسازی الگوهای ژنتیکی پنهان میان تیره‌ها کمک کرد.



شکل ۲. نمودار ارزش تجمعی مقادیر مولفه های اصلی

Figure 2. Cumulative value diagram of principal component values

با توجه به این که در این مطالعه از روش های یادگیری ماشین جهت طبقه بندی تیره ها استفاده شد جهت بهبود روند لازم بود داده های پرتی که باعث خطا در آنالیز ها میشوند را حذف کرد که در نتیجه تعداد ۷ نمونه از تیره خراسان و کل نمونه های شویمان به دلیل تعداد کم نمونه ها از آنالیز های یادگیری ماشین کنار گذاشته شد و به تعداد نمونه ها به ۹۹ راس اسب عرب ایران (اصیل) شامل (خرسان: ۴۶، حمدانی: ۱۹، عبیان: ۱۳، صغلاوی: ۱۲، کحیلان: ۹) رسید. به جهت مقایسه نتایج آنالیز های انجام شده از ماتریس های درهم ریختگی (Confusion Matrix) استفاده شد. این ماتریس نتایج را برای جفت های مختلف جمعیت ها نشان می دهند که با استفاده از اطلاعات یک مدل طبقه بندی، ماتریسی رسم میکند که نشان دهنده عملکرد مدل در تشخیص دو جمعیت خاص است. این ماتریس برچسب های پیش بینی شده توسط مدل را نشان می دهند و برچسب های واقعی (پیش بینی درست) را نشان می دهند، که اعداد مربوط به ماتریس درهم ریختگی به صورت جدول نمایش داده شد (جدول ۱) اعداد داخل جدول تعداد نمونه هایی را نشان می دهند که در یک دسته خاص به درستی یا نادرستی پیش بینی شده اند. در مدل RF مربوط به نتایج ماتریس درهم ریختگی، تیره های "خرسان" و "عبیان"، عدد ۱۴ در کلاس بندی درست (کلاس ۱ درست) (که به این معنی است، ۱۴ نمونه از جمعیت "خرسان" به درستی به عنوان "خرسان" و عدد ۴ در (کلاس ۲ درست) به معنی ۴ نمونه از تیره "عبیان"، درست پیش بینی شده اند (جدول ۱). اگر عددی در بخش کلاس ۱ یا ۲ اشتباه قرار گیرد (یعنی جایی که پیش بینی شده با واقعیت مطابقت ندارد)، نشان دهنده خطای مدل است. به عنوان مثال، در مقایسه تیره های "خرسان" و "عبیان"، الگوریتم RF, SVM موفق به شناسایی تمامی نمونه های "خرسان" و ۴ نمونه از "عبیان" با دقت پیش بینی ۱۰۰ درصد گردید و دقت پیش بینی الگوریتم MLP ۸۲ درصد شد که نشان دهنده عملکرد ضعیف این الگوریتم بود، نتایج این مقایسه نشان

دهنده تمایز ژنتیکی واضح بین این دو تیره است (جدول ۱). در برخی موارد، مدل ها در تفکیک تژادهایی که از نظر ژنتیکی شباهت بیشتری دارند، دقت کمتری از خود نشان داد. به عنوان مثال، در مقایسه "صغلاوی" با "عبیان"، میزان نمونه های اشتباه طبقه بندی شده در جدول نتایج ماتریس درهم ریختگی (جدول ۱) افزایش یافت به طوری که در هر سه الگوریتم RF, SVM, MLP, به ترتیب ۵۲، ۵۳ و ۴۷ درصد دقت پیشبینی برآورد شد که می تواند نشانه ای از نزدیکی ژنتیکی این دو نژاد باشد. در بخش طبقه بندی مربوط به روش SVM در بخش نتایج ماتریس های درهم ریختگی نشان داد که اکثر جفت ها مانند تیره های (خرسان، کوهیلان) (حمدانی خرسان) (خرسان صغلاوی) (خرسان عبیان) (حمدانی کوهیلان) با دقتی بالای ۸۰ درصد و خطا های کمتر نسبت به RF, MLP در تشخیص متقابل تقریباً به طور کامل از هم تفکیک می شوند (جدول ۱). در برخی از جفت ها خطای مدل مشهود است به طوری که دقت پیشبینی مدل SVM در مقایسه جفت تیره ها مانند (حمدانی، صغلاوی) (صغلاوی، عبیان) به ترتیب ۵۹ و ۵۳ درصد شد که باز هم از دو مدل مورد استفاده در این مطالعه بهتر بود که در بررسی نتایج ماتریس درهم ریختگی نشان داد که تیره های (صغلاوی، عبیان) (حمدانی صغلاوی) الگوی نزدیکی بیشتری از لحاظ ژنتیکی به یکدیگر داشتند. نتایج ماتریس درهم ریختگی (جدول ۱) در تمام زوج جمعیت ها حاکی از عملکرد کلی مطلوب الگوریتم SVM در تفکیک تژادها بود، اما وجود نمونه های اشتباه طبقه بندی شده در برخی زوج ها، تأکید بر لزوم استفاده از ویژگی ها یا الگوریتم های تکمیلی برای بهبود دقت مدل دارد. برای مثال، بین دو تیره صغلاوی و حمدانی بیشتر نمونه ها اشتباهاً به جای یکدیگر طبقه بندی شده اند و این نشان دهنده شباهت های مورفولوژیک یا هم پوشانی ویژگی های ژنتیکی در این دو تیره است (جدول ۱). این نتایج نشان می دهد با وجود تفکیک پذیری کلی، هنوز نقاط ابهام قابل توجهی در تفکیک تیره های اسب عرب ایران (اصیل) وجود دارد. به طور کلی، نتایج ما بیانگر آن است که SVM با کرنل خطی برای اکثر جفت ها تفکیک قابل قبولی ارائه می کند، اما در مواردی که شکل یا ویژگی های مشترک بیشتری بین تیره ها وجود دارد، نیاز به افزودن متغیرهای توصیفی یا بررسی هسته های غیرخطی احساس می شود. ارزیابی عملکرد مدل شبکه عصبی چندلایه MLP توانست، با استفاده از یک لایه مخفی ۱۰ نورونی داده ها را تفکیک کند و کلاس بندی قابل قبولی در حد RF ارائه دهد. همان طور که در نتایج ماتریس درهم ریختگی (جدول ۱) مشخص است، دقت مدل SVM در تفکیک کلاس ها متغیر بوده و بیشترین چالش مربوط به شباهت بین کلاس هایی مانند (صغلاوی، عبیان) (کوهیلان، عبیان) یا (حمدانی، صغلاوی) بوده است. با این حال، بهبود عملکرد مدل در برخی از ترکیب های کلاس، به خصوص در مواردی که داده های تفکیک شده واضح تر هستند، نشانگر کارایی مناسب SVM در یادگیری ویژگی های کلیدی داده ها بوده است. تعداد نمونه ها، تنظیمات مدل و تعداد تکرارهای بالا، نقش مهمی در همگرایی و کاهش خطاهای دسته بندی داشته اند. جهت ارزیابی مدل ها بر اساس میزان معنی داری تفاوت بین مدل ها آزمون فریدمن انجام شد و مقدار آماره آزمون test statistic معادل ۱۷.۰۵ و مقدار p-value برابر با ۰.۰۰۰۱۹ است که تفاوت معنی داری بین مدل ها را آشکار کرد و در نتیجه تفاوت معنی داری بین مدل ها وجود دارد و انتخاب مدل با توجه به نتیجه آزمون فریدمن جهت انجام مطالعات پیشرو

اهمیت دارد. نتیجه عملکرد مدل‌ها (جدول ۲) نشان می‌دهد که مدل SVM با میانگین دقت ۰.۸۳ نسبت به دو مدل RF با دقت ۰.۷۵ و MLP با دقت ۰.۷۳ بالا ترین عملکرد را نشان داد و مدل MLP دقت و پایداری کمتری را نسبت به دیگر مدل‌ها دارد و مدل RF دقت کمی بهتر از MLP ارائه کرد. در مطالعه ای جهت تفکیک جمعیت‌های انسانی از روش‌های یادگیری ماشین تحت نظارت، ماشین بردار پشتیبان و شبکه عصبی استفاده شد، که SVM و شبکه عصبی به طور قابل توجهی در تفکیک و جداسازی موفق بودند (Bridges *et al.*, 2011). نتایج مطالعات (Panahiazar *et al.*, 2015; Fernández-Delgado *et al.*, 2014) حاکی از آن است که شبکه‌های عصبی چندلایه می‌توانند در داده‌های پیچیده و مسائل چندکلاسه عملکرد برتری نسبت به برخی مدل‌های سنتی داشته باشد و در مطالعه دیگری مشخص شد، در مطالعات ژنتیکی مرتبط با شناسایی نژادهای دامی، تحقیقات نشان داده‌اند که RF در شناسایی تنوع ژنتیکی و ترکیب ژنوتیپی ساده‌تر عملکرد مناسبی دارد، اما وقتی صفات مورد بررسی پیچیده‌تر می‌شوند، روش‌های مبتنی بر شبکه‌های عصبی مصنوعی مانند MLP نتایج دقیق‌تری ارائه داده‌اند (Mrode *et al.*, 2020).

جدول ۱. تفکیک زوجی تیره‌ها با استفاده از روش‌های یادگیری ماشین (RF, SVM, MLP)
Table 1. Pairwise separation of families using machine learning methods (RF, SVM, MLP)

زوج تیره‌ها	مدل	پیش‌بینی مدل برای کلاس ۱	کلاس ۱ درست	کلاس ۱ اشتباه	پیش‌بینی مدل برای کلاس ۲	کلاس ۲ درست	کلاس ۲ اشتباه
Khersan vs Koheilan	random forest		14	0		2.7	0.3
	svm	Khersan	14	0	Koheilan	2.9	0.1
	mlp		12	2		3	0
Hamdani vs Khersan	random forest		0.8	5.2		14	0
	svm	Hamdani	3.4	2.6	Khersan	14	0
	mlp		4.7	1.3		13	0.6
Khersan vs Saghlavi	random forest		14	0		0.3	3.7
	svm	Khersan	14	0	Saghlavi	3	1
	mlp		13	1.4		3.8	0.2
Hamdani vs Koheilan	random forest		5.8	0.2		2.3	0.7
	svm	Hamdani	6	0	Koheilan	2.6	0.4
	mlp		4.1	1.9		2.5	0.5
Khersan vs Obian	random forest		14	0		4	0
	svm	Khersan	14	0	Obian	4	0
	mlp		11	2.6		3.9	0.1
Obian vs Saghlavi	random forest		3.6	0.4		0.2	3.8
	svm	Obian	3.5	0.5	Saghlavi	0.8	3.2
	mlp		1.2	2.88		3	1
Hamdani vs Saghlavi	random forest		4.5	1.5		0.9	3.1
	svm	Hamdani	5.6	0.4	Saghlavi	0.3	3.7
	mlp		3	3		2.5	1.5
Hamdani vs Obian	random forest		5.6	0.4		1.9	2.1
	svm	Hamdani	5.8	0.2	Obian	2.4	1.6
	mlp		3.5	2.5		3.3	0.7
Koheilan vs Saghlavi	random forest		2.5	0.5		2.4	1.6
	svm	Koheilan	2.7	0.3	Saghlavi	3.3	0.7
	mlp		1.7	1.3		3.3	0.7
Koheilan vs Obian	random forest		1.3	1.7		3.2	0.8
	svm	Koheilan	1.8	1.2	Obian	3.5	0.5
	mlp		1.3	1.7		3.2	0.8

در مجموع و با توجه به نتایج آزمون فرید من، می‌توان نتیجه گرفت که انتخاب مدل SVM برای داده‌های این مطالعه گزینه‌ای بهینه‌تر است و همچنین بررسی نتایج ماتریس‌های درهم ریختگی نشان داد که SVM در تفکیک کلاس‌ها موفق‌تر از سایر مدل‌ها بوده است. این نتایج نشان می‌دهد، هنگامی که با کمبود داده مواجه هستیم مدل SVM از شبکه‌ی عصبی MLP بهتر

عمل میکند. در کل انتخاب مدل باید متناسب با ویژگی های داده و هدف مطالعه باشد، اما SVM با داده های پیچیده ژنتیکی عملکرد برتری از خود نشان داده است. همچنین، مدل MLP از نظر کارایی، انحراف معیار کمتری نسبت به SVM و RF نشان داد. این نتایج بیانگر عملکرد برتر مدل SVM برای طبقه بندی ژنتیکی تیره های اسب عرب ایرانی در این مطالعه است.

جدول ۲. نتایج میانگین دقت آزمون فریدمن

Table 2. Average accuracy results of the Friedman test

مدل	میانگین دقت	انحراف معیار
RF	0.752	0.175
SVM	0.833	0.162
MLP	0.739	0.144

نتیجه گیری

به طور کلی، نتایج حاصل از کاربرد این مدل ها در مطالعات ژنتیک جمعیت، چه دامی و چه انسانی، تأیید می کند که SVM در داده های با ویژگی های تفکیک پذیر مناسب تر عمل می کند. از سوی دیگر، MLP در شرایطی که روابط بین داده ها پیچیده تر است، دقت بالاتری دارد اما به دلیل وجود دیتای کم در مقایسه با دو مدل دیگر مخصوصاً در تفکیک داده های پیچیده ژنتیکی در رتبه پایین تری قرار گرفته است. مدل RF بیشتر در پیش بینی های ساده تر و داده های کاملاً مشهود کارایی دارد و در مقایسه با مدل SVM و در تفکیک داده های پیچیده ژنتیکی در رتبه پایین تری قرار گرفته است.

تشکر و قدردانی

در پایان لازم است از شرکت تعاونی میثاق همدان با حمایت های بیدریغ و آقای مهندس امیر حسین حیدری جهت یاری در پیش برد این مقاله در آزمایشگاه بیوانفورماتیک دانشگاه تهران، کمال تقدیر و تشکر به عمل آید.

منابع

Analytics Steps. (n.d.). Applications of Machine Learning in Genetics and Genomics. Analytics Steps. Retrieved from <https://analyticssteps.com/blogs/10-applications-machine-learning-genomics>

Badbarin, S., Sharifi, R.S. and Falahi, H., (2022). Investigation the genetic diversity of Arabian horses and their different strains using SSR markers.

Bashardoost, M. (2020). Design of SNP chip for identification of Turkmen and Kurdish horse breeds using 70K SNP_Chip genomic data. MSc thesis, University of Tehran, College of Natural Resources, Karaj. In Persian

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Bridges, M., Heron, E. A., O'Dushlaine, C., Seguro, R., et al. (2011). Genetic classification of populations using supervised learning. *PLoS ONE*, 6(5), e14802. <https://doi.org/10.1371/journal.pone.0014802>

- Cervantes, I., Gutiérrez, J. P., Valera, M., Goyache, F., & Molina, A. (2020).** Machine learning for breed classification using single nucleotide polymorphisms in domestic animals. **Frontiers in Genetics*, 11*, 573.
- Cortes, C., & Vapnik, V. (1995).** Support-vector networks. *Machine Learning*, 20(3), 273-297.
- Fernandez-Delgado, M., Cernadas, E., Barro, S & Amorim, D. (2014).** Do we need hundreds of classifiers to solve real world classification problems ? *Journal of Machine Learning Research*, ۱۵, ۳۱۳۳-۳۱۸۱.
- Forbis, J. A. (1986).** *The classic Arabian horse*. Howell Book House.
- Ghasemi, E., Mehrabani-Yeganeh, H., Ajmone-Marsan, P., Razmkabir, M., Naghavi, H. R., ... & Moradi, M. H. (2021).** Deciphering genomic diversity and selection signatures in Iranian native horses using whole-genome sequencing. *Frontiers in Genetics*, 12, 665382.
- Głazewska, I., (2010).** Speculations on the origin of the Arabian horse breed. *Livestock Science*, 129(1-3), pp.49-55.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009).** The elements of statistical learning: data mining, inference, and prediction. **Springer Science & Business Media.**
- Hill, E. W., McGivney, B. A., Gu, J., Whitten, M., MacHugh, D. E., & Katz, L. M. (2012).** Inferring evolutionary relationships among modern horse breeds using dense genome-wide SNP data. *PloS Genetics*, 8(12), e1002451.
- Hill, E. W., McGivney, B. A., Gu, J., Whitten, M., MacHugh, D. E & Katz, L. M. (2021).** Recent advances in understanding the genomic basis of equine performance. *Frontiers in Genetics*. ۱۲, ۷۵۳۰۲۳.
- Khaleili, M. (2008).** The horse and what I know about it. Tehran: Zarre Publishing.
- Khanshour, A. (2013).** Population structure and inbreeding of Arabian horse populations. Master's thesis, University of Florida.
- Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993).** Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6), 861-867.
- Liaw, A., & Wiener, M. (2002).** Classification and regression by random forest. *R News*, 2(3), 18–22.
- Liu, R., Xu, Z., Teng, J., Pan, X., Lin, Q., Cai, X., et al. (2022).** Evaluation of six machine learning classification algorithms in pig breed identification using SNPs array data. *Animal Genetics*, 54(1), 1-10.
- Motsinger-Reif, A., & Ritchie, M. D. (2008).** Neural networks in genetic epidemiology; past, present and future. *BioData Mining*, 1, 3.
- Mrode, R. A., Vahedi, H., Allahghadri, M., Ghorbani, G. R & Sadeghi, S. (2020).** Application of artificial neural networks and support vector machines in genomic prediction of milk yield and milk components in Iranian Holstein dairy cattle. *Animal*. ۹۸۳-۹۹۲, (۵)۱۴.
- Ogutu, J. O., Piepho, H. P., & Schulz-Streeck, T. (2011).** A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, 5(Suppl 3), S11.
- Panahiazar, M., Taslimitehrani, V., Jadhav, A & Pathak, J. (2015).** Predicting risk of emergency department visits using a machine learning algorithm. *BMC Medical Informatics and Decision Making*, ۱۵(۱), ۱۴۰.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011).** Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Daly, M. J. (2007).** PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3): 559-575.
- Sadeghi, R., Moradi-Shahrbabak, M., Miraei Ashtiani, S.R., Schlamp, F., Cosgrove, E.J. and Antczak, D.F., (2019).** Genetic diversity of Persian Arabian horses and their relationship to other native Iranian horse breeds. *Journal of Heredity*, 110(2), pp.173-182.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Valdes, J. M. P., Soto, D. P., Bussiman, F., & Oliveira, T. F. (2023). Detecting SNP markers discriminating horse breeds by machine learning techniques. *BMC Bioinformatics, 24*(1), 290