**Animal Production Research**

Vol. 14, No. 1, 2025 (1-17)
doi: 10.22124/ar.2025.29391.1872
eISSN: 2538-6107 pISSN: 2252-0872

University of Guilan

**RESEARCH PAPER** **OPEN ACCESS**

# Evaluating the performance of support vector machine with different kernels in genomic analysis at different levels of dominance variance

**H. Sahebalam, M. Gholizadeh*, H. Hafezian**

Department of Animal Science, Faculty of Animal Science and Fisheries, Sari Agricultural Sciences and Natural Resources University, Sari, Iran

(Received: 26-12-2024 – Revised: 26-02-2025 – Accepted: 27-02-2025 – Available online: 28-03-2025)

**Abstract**

**Introduction:** Predicting quantitative traits is a fundamental aspect of plant and animal breeding. Genomic selection is a precise and efficient approach that estimates genetic merit using high-density single nucleotide polymorphisms (SNPs). However, most genomic selection procedures primarily focus on additive effects to calculate the genomic estimated breeding value (GEBV) for selection candidates. Nonetheless, incorporating non-additive effects offers several advantages: (i) it enhances the accuracy of GEBV predictions and subsequent selection responses, (ii) it facilitates optimized mate allocation among selection candidates, and (iii) it enables improved utilization of non-additive genetic variation through tailored crossbreeding or purebred breeding strategies. One of the most challenging factors affecting the accuracy of genomic evaluation is the selection of an appropriate statistical method to estimate marker effects with high accuracy. The most common parametric methods for genomic evaluation are the genomic best linear unbiased predictor (GBLUP) and Bayesian methods, which use the co-variance structure between individuals and regression of phenotype on markers to predict the genetic values of individuals, respectively. However, in recent years, non-parametric methods of machine learning have been widely used for genomic evaluation in animal and plant breeding programs. This study aimed to evaluate and compare the accuracy of genomic predictions using GBLUP and support vector machine (SVM) methods. The SVM models employed various kernel functions, including linear (SVM-lin), Gaussian radial (SVM-rad), polynomial (SVM-pol), and cyclic (SVM-sig). Both purely additive and additive + dominance deviation gene action models were considered under varying levels of dominance variance.

**Materials and methods:** A simulated genome comprising six chromosomes with a total length of 600 cM was used for this study. Each chromosome contained 1,000 evenly spaced SNPs and 100 randomly distributed quantitative trait loci (QTLs). Phenotypic variance ($\sigma_p^2$) and narrow-sense heritability ($h^2$) were set to 1 and 0.4, respectively. Dominance variance ($\sigma_d^2$) levels were evaluated at 0.10, 0.15, 0.20, 0.25, 0.30, and 0.35. Prediction accuracy was calculated as the Pearson correlation coefficient between the true genetic value (TGV) or true breeding value (TBV) and their respective genomic estimates (GEGV or GEBV). The correlations were represented as r(TGV, GEGV) and r(TBV, GEBV), respectively. In addition, the practical significance of differences in prediction accuracy among the studied statistical methods was assessed using Cohen's d effect size during 100 replicates.

**Results and discussion:** The conventional GBLUP method consistently exhibited higher prediction accuracy for both GEBV and GEGV across all scenarios. Among the SVM approaches, the SVM-rad and SVM-sig kernels showed superior performance in predicting GEGV under both purely additive and additive + dominance deviation models. However, for GEBV prediction, their performance declined with increasing dominance variance. When

* Corresponding author: m.gholizadeh@sanru.ac.ir

CC BY

dominance variance exceeded 0.30, SVM-lin and SVM-sig demonstrated prediction accuracy comparable to or slightly better than SVM-rad. The differences in prediction accuracy between GBLUP and SVM-rad were minimal (d=0.218) in the purely additive model but reached their peak (d=0.492 and d=0.404) in the additive + dominance deviation model at the highest dominance variance ($\sigma_d^2$=0.35). This disparity occurred because, as dominance variance increased, the GBLUP method exhibited slightly greater changes in accuracy compared to the SVM-rad method. Furthermore, with higher dominance variance, the difference in prediction accuracy for GEBV between the SVM methods and both GBLUP and SVM-rad substantially decreased. For example, in the purely additive model ($\sigma_d^2$=0), Cohen's d was 2.608 and 2.336, respectively, while in the additive + dominance deviation model ($\sigma_d^2$=0.35), d dropped to 0.309 and 0.189, respectively. Using the additive + dominance deviation model significantly improved GEBV prediction accuracy, particularly when dominance variance contributed substantially to phenotypic variance. This improvement is due to dominance deviation, which arises from interactions between alleles at a locus. While additive effects are represented as breeding values, which partially incorporate dominance effects, genomic evaluations that explicitly consider dominance effects can further enhance GEBV accuracy.

**Conclusions:** The choice of kernel function in SVM models plays a pivotal role in the accuracy of GEBV and GEGV predictions. Overall, when applying the nonparametric SVM method to fit markers to phenotypes, the Gaussian radial kernel function is recommended for optimal performance. However, as the dominance variance increased, the performance of SVM-lin and SVM-sig methods improved significantly and the performance gap with GBLUP and SVM-rad decreased. This indicated the potential capacity of SVM in investigating non-additive effects, especially in situations where the contribution of dominance in explaining phenotypic variance increases.

**Keywords:** Genomic breeding value, Kernel function, Genomic analysis, Prediction accuracy, Support vector machine

**Ethics statement:** This article does not contain any studies with human participants or animals performed by any of the authors.

**Data availability statement:** The data that support the findings of this study are available on request from the corresponding author.

**Conflicts of interest:** The authors declare no conflicts of interest.

**Funding:** The authors received no specific funding for this project.

**How to cite this article:**

Sahebalam, H., Gholizadeh, M., & Hafezian, H. (2025). Evaluating the performance of support vector machine with different kernels in genomic analysis at different levels of dominance variance. *Animal Production Research*, *14*(1), 1-17. doi: 10.22124/ar.2025.29391.1872

تحقیقات تولیدات دامی

سال چهاردهم/شماره اول/بهار ۱۴۰۴ (۱۷-۱)

**مقاله پژوهشی**

# ارزیابی عملکرد ماشین بردار پشتیبان با کرنل‌های مختلف در تجزیه ژنومی در سطوح مختلف واریانس غالبیت

**حمید صاحب علم، محسن قلی زاده*، حسن حافظیان**

گروه علوم دامی، دانشکده علوم دامی و شیلات، دانشگاه علوم کشاورزی و منابع طبیعی ساری

(تاریخ دریافت: ۱۴۰۳/۱۰/۰۶ – تاریخ بازنگری: ۱۴۰۳/۱۲/۰۸ – تاریخ پذیرش: ۱۴۰۳/۱۲/۰۹ – تاریخ انتشار برخط: ۱۴۰۴/۰۱/۰۸)

## چکیده

هدف از پژوهش حاضر، بررسی و مقایسه صحت پیش‌بینی ژنومی روش ماشین بردار پشتیبان (SVM) بر اساس توابع کرنل مختلف شامل خطی (SVM-lin)، شعاعی (SVM-rad)، چندجمله‌ای (SVM-pol) و حلقوی (SVM-sig)، و روش GBLUP در مدل‌های کنش ژنی صرفاً افزایشی و افزایشی-انحراف غالبیت با در نظر گرفتن سطوح مختلف واریانس غالبیت بود. بدین منظور، ژنومی حاوی شش کروموزوم و به‌طول ۶۰۰ سانتی‌مورگان شبیه‌سازی شد. روی هر کروموزوم، ۱۰۰۰ نشانگر چندشکلی تک نوکلئوتیدی (SNP) با فواصل یکسان و ۱۰۰ جایگاه صفت کمّی (QTL) به‌طور تصادفی در نظر گرفته شد. واریانس فنوتیپی و وراثت پذیری به‌ترتیب برابر با ۱ و ۰/۴ در نظر گرفته شد. واریانس انحراف غالبیت برابر با ۰/۱۰، ۰/۱۵، ۰/۲۰، ۰/۲۵، ۰/۳۰ و ۰/۳۵ در نظر گرفته شد. صحت پیش‌بینی به‌عنوان ضریب همبستگی پیرسون بین ارزش ژنتیکی واقعی (TGV) یا ارزش اصلاحی واقعی(TBV) و ارزش ژنتیکی ژنومی (GEGV) یا ارزش اصلاحی ژنومی (GEBV) تعریف شد. روش مرسوم GBLUP در تمام سناریوهای مختلف واریانس غالبیت، صحت پیش‌بینی GEBV و GEGV بالاتری را نشان داد. در بین رویکردهای مختلف SVM، در مدل صرفاً افزایشی و افزایشی- انحراف غالبیت بر اساس صحت پیش‌بینی GEGV، رویکردهای SVM-rad و SVM-sig به‌ترتیب بالاترین عملکرد را نشان دادند. بر اساس صحت پیش‌بینی GEBV، با افزایش واریانس غالبیت، این برتری به‌شدت کاهش یافت، به‌طوری که در واریانس غالبیت بیشتر از ۰/۳۰، رویکردهای SVM-lin و SVM-sig به‌ترتیب صحت پیش‌بینی GEBV اندکی بالاتر و برابر با SVM-rad نشان دادند. به‌طورکلی، در برازش فنوتیپ روی نشانگرها با روش ناپارامتری SVM، استفاده از تابع کرنل شعاعی در مدل پیشنهاد می‌شود.

**واژه‌های کلیدی:** ارزش اصلاحی ژنومی، تابع کرنل، تجزیه ژنومی، صحت پیش‌بینی، ماشین بردار پشتیبان

* نویسندهٔ مسئول: m.gholizadeh@sanru.ac.ir

doi: 10.22124/ar.2025.29391.1872

**مقدمه**

پیش از عصر ژنومی (از دهه ۱۹۷۰ تا اوایل دهه ۲۰۰۰)، تلاش‌ها برای اصلاح نژاد دام از راه انتخاب و توسعه معادلات مدل مختلط برای بهترین پیش‌بینی کننده نااریب خطی (BLUP)، ساخت معکوس ماتریس روابط خویشاوندی شجره (Quaas, 1976; Henderson, 1976) و برآورد فراسنجه‌های ژنتیکی صفات اقتصادی بود. توسعه نشانگرهای مولکولی در دهه‌های اخیر، فرصتی را برای دسترسی و استفاده از نشانگرهای چندشکلی بسیار متراکم در اصلاح حیوانات و گیاهان فراهم کرده است (Crossa *et al.,* 2017). چندشکلی‌های تک نوکلئوتیدی (SNPs) کاربردی‌ترین سامانه تعیین ژنوتیپ با توان عملیاتی بالا هستند که به‌طور گسترده در شناسایی مکان‌های صفت کمّی (QTL) استفاده شده‌اند. در انتخاب ژنومی، فرض بر این است که تمام واریانس ژنتیکی صفت به‌وسیله نشانگرها توجیه می‌شود، به‌طوری که هر QTL با حداقل یک SNP در عدم تعادل پیوستگی است. علاوه بر این، آثار همه نشانگرها به‌طور همزمان برآورد می‌شوند. انتخاب ژنومی داده‌های مولکولی و فنوتیپی را در یک جمعیت مرجع ترکیب می‌کند تا آثار نشانگری را برای به‌دست آوردن GEBVهای افراد جمعیت هدف که دارای ژنوتیپ و فاقد فنوتیپ هستند، پیش‌بینی کند (Meuwissen *et al.,* 2001). از آنجا که شایستگی ژنتیکی افزایشی افراد به‌طور مستقیم به نسل بعدی منتقل می‌شود، آثار غیرافزایشی آلل‌ها در ارزیابی‌های ژنتیکی، به شکل مرسوم، نادیده گرفته می‌شوند (Varona *et al.,* 2018). غالبیت، به‌عنوان منبع واریانس ژنتیکی غیرافزایشی، برهمکنش بین آلل‌ها در یک مکان کروموزومی است. هنگامی که آثار ژنتیکی غیر-افزایشی منبع مهمی از تنوع ژنتیکی باشد، استفاده از آثار افزایشی یا ارزش اصلاحی (BV)، به تنهایی، ممکن است منجر به انتخاب ژنوتیپ‌هایی شود که بالاترین ارزش ژنتیکی را ندارند. اطلاعات ژنومی می‌تواند با برآورد ارزش ژنتیکی یک فرد که منجر به پیش‌بینی‌های دقیق‌تر برای فنوتیپ‌های آینده شود، امکان برآورد آثار ژنتیکی غیر-افزایشی نشانگرها را فراهم کند (Toro and Varona, 2010; Vitezica *et al.,* 2013). برآورد آثار ژنتیکی غیرافزایشی می‌تواند در افزایش صحت پیش‌بینی ارزش‌های اصلاحی و پاسخ به انتخاب نقش داشته باشد (Aliloo *et al.,* 2016; Duenk *et al.,* 2017; Toro and Varona, 2010). این آثار، تعریف و تعیین روش‌های آمیزش بین افراد کاندیدا را تسهیل می‌کنند (Maki-Tanila, 2007; Toro and Varona, 2010; Aliloo *et al.,* 2017). همچنین، این آثار می‌توانند برای بهره‌گیری از تنوع ژنتیکی غیرافزایشی در برنامه‌های اصلاح نژاد از راه آمیخته‌گری و تلاقی لاین خالص استفاده شوند (Maki-Tanila, 2007; Zeng *et al.,* 2013). معمولاً در مدل‌های ارزیابی ژنتیکی سنتی، آثار غیرافزایشی به دلایل مختلفی از جمله عدم وجود شجره‌نامه‌های کامل و دقیق و نیاز به محاسبات پیچیده‌تر، کمتر مورد توجه قرار می‌گیرند (Varona *et al.,* 2018). پژوهش‌های مختلف رویکردهای متفاوتی را برای بکارگیری آثار غالبیت در مدل‌های ژنومی پیشنهاد نموده‌اند (Toro and Varona, 2010; Su *et al.,* 2012; Vitezica *et al.,* 2013).

شبیه‌سازی داده‌های ژنومی، فرصتی مناسب در راستای مطالعه اثر عوامل مختلف بر صحت پیش‌بینی ژنومی با هزینه بسیار پایین را ممکن می‌سازد (Thomasen *et al.,* 2013)، که به‌طورگسترده در اصلاح نژاد دام مورد استفاده قرار گرفته است (Saheb Alam *et al.,* 2018; Atefi *et al.,* 2021; Tamaddoni-Arani *et al.,* 2021; Ansari *et al.*, 2024). یکی از چالش برانگیزترین عوامل موثر بر صحت ارزیابی ژنومی، انتخاب روش آماری مناسب برای برآورد آثار نشانگری، با صحت بالا است. بسیاری از روش‌ها از جمله روش‌های پارامتری و ناپارامتری، با طیف گسترده‌ای از صحت پیش‌بینی برای ارزیابی ژنومی پیشنهاد شده‌اند (Meuwissen *et al.,* 2001; Ogutu *et al.,* 2011; Howard *et al.,* 2014; Akbarpour *et al.*, 2021; Sahebalam *et al.,* 2024). رایج‌ترین روش‌های پارامتری در ارزیابی ژنومی، روش‌های بهترین پیش‌بینی کننده نااریب خطی ژنومی (GBLUP) و بیزی هستند که بترتیب از ساختار واریانس-کوواریانس بین افراد و رگرسیون مستقیم فنوتیپ بر نشانگرها برای پیش‌بینی ارزش‌های ژنتیکی افراد استفاده می‌کنند (de los Campos *et al.,* 2009). با این حال، در سال‌های اخیر از روش‌های ناپارامتری یادگیری ماشین به‌طور گسترده‌ای برای ارزیابی ژنومی در برنامه‌های اصلاح نژاد حیوانات و گیاهان استفاده شده است (Ghafouri-Kesbi *et al.,* 2017; Sahebalam *et al.,* 2019). ماشین بردار پشتیبان یکی از روش‌های یادگیری ماشین است (Boser *et al.,* 1992) که علاوه بر عملکرد مطلوب در ارزیابی ژنومی، در تشخیص برهمکنش

پروتئین-پروتئین، برهمکنش ژن-محیط، شناسایی عوامل تنظیمی در توالی آمینواسید، ژن‌های مرتبط با بیماری، مدل‌سازی جهت ارتباط میان ترکیب نشانگرها و شناسایی ژن‌های هدف، استفاده می‌شود (Yang *et al.,* 2010). در پژوهش‌های مختلف بر اساس داده شبیه‌سازی، عملکرد پیش‌بینی ماشین بردار پشتیبان با سایر روش‌های پارامتری و ناپارامتری مورد مقایسه قرار گرفته است (Ogutu *et al.,* 2011; Sahebalam *et al.,* 2019). در مدل ماشین بردار پشتیبان، ساخت مدل شامل دو مرحله یادگیری و اعتبارسنجی است. در انتهای فاز یادگیری، قابلیت تعمیم مدل یادگیری با استفاده از داده‌های اعتبارسنجی ارزیابی می‌شود. ماشین بردار پشتیبان الگوریتمی است که نوع خاصی از مدل خطی را پیدا می‌کند که حاشیه ابر-صفحه (hyper-plane margin) را به حداکثر می‌رساند. به حداکثر رساندن حاشیه ابر-صفحه منجر به حداکثر رساندن جدایی بین لایه‌ها می‌شود، و به نزدیکترین نقاط آموزشی به این حاشیه، بردارهای پشتیبان اطلاق می‌شود. فقط از این بردارها یا نقاط برای تعیین مرز بین لایه‌ها استفاده می‌شود (Shin *et al.,* 2005). اگر داده‌ها به‌شکل خطی از یکدیگر جدا شوند، ماشین بردار پشتیبان، ماشین‌های خطی را برای تولید یک سطح بهینه که داده‌ها را بدون خطا و با حداکثر فاصله بین صفحه و نزدیکترین نقاط یادگیری (بردارهای پشتیبان) جدا می‌کند، آموزش می‌دهد. اگر داده‌ها به‌صورت خطی قابل تفکیک نباشند، ماشین بردار پشتیبان، برای ایجاد ماشین‌هایی با انواع مختلفی از سطوح تصمیم‌گیری غیرخطی در فضای داده‌ها، از تابع کرنل استفاده می‌کند. تابع کرنل، یک تابع فضای اولیه است، که برابر با ضرب داخلی دو بردار در فضای ویژگی (feature space) است. تابع کرنل باید یک تابع مثبت معین متقارن باشد تا با ضرب داخلی دو بردار در فضای ویژگی، معادل شود. تابع کرنل از دو ورودی شامل x (بردار ژنوتیپ برای هر فرد) و y (ارزش-های فنوتیپی) استفاده می‌کند و به الگوریتمی مانند ماشین بردار پشتیبان اجازه می‌دهد تا آنها را به آسانی پردازش کند (Hastie *et al.,* 2009). توابع کرنل شامل خطی، چند-جمله‌ای، شعاعی گاوسی و حلقوی هستند. در پژوهشی با-استفاده از داده‌های شبیه‌سازی شده و با در نظر گرفتن تنها آثار افزایشی برای QTLها، عملکرد پیش‌بینی روش ماشین بردار پشتیبان مبتنی بر کرنل‌های خطی، چندجمله‌ای، شعاعی و حلقوی برای ارزیابی ژنومی صفات آستانه‌دار مورد بررسی قرار گرفت (Kasnavi *et al.,* 2018). هدف این مطالعه، مقایسه عملکرد پیش‌بینی ژنومی روش‌های GBLUP و ماشین بردار پشتیبان با استفاده از توابع کرنل مختلف (خطی، شعاعی گاوسی، چندجمله‌ای و حلقوی) در مدل‌های کنش ژنی افزایشی و افزایشی-انحراف غالبیت با در نظر گرفتن سطوح مختلف واریانس غالبیت بود.

## مواد و روش‌ها

*شبیه‌سازی داده‌ها:* جمعیت‌های مورد استفاده در این پژوهش با استفاده از بسته نرم افزاری xbreed (Esfandiari and Sorensen, 2017) در محیط نرم افزار R شبیه‌سازی شد. به‌منظور ایجاد عدم تعادل پیوستگی (LD) و ایجاد تعادل جهش-رانش، جمعیت تاریخی با دو نیروی جهش و رانش شبیه‌سازی شد. جمعیت تاریخی شامل ۱۰۰۰ نسل آمیزش تصادفی با اندازه موثر ۳۰۰ فرد (۱۵۰ نر و ۱۵۰ ماده) بود. نتاج از ترکیب تصادفی گامت‌های پدری و مادری ایجاد شدند. اندازه جمعیت با تعداد مساوی نر و ماده در طول نسل‌های گسسته ثابت ماند. آخرین نسل از جمعیت تاریخی به‌عنوان نسل پایه‌گذار جمعیت اخیر در نظر گرفته شد. در نسل پایه جمعیت اخیر، تعداد ۲۰۰ فرد (۱۰۰ نر و ۱۰۰ ماده) به‌طور تصادفی از آخرین نسل تاریخی (نسل ۱۰۰۰) انتخاب شدند. این جمعیت تا سه نسل ادامه یافت، به‌طوری که والدین هر نسل از نسل قبل انتخاب شدند. در این جمعیت، تعداد نتاج به ازای هر مادر، پنج فرد در نظر گرفته شد. در نسل اول جمعیت اخیر، تعداد افراد به ۵۰۰ فرد افزایش یافت. ۴۰۰ فرد (۲۰۰ نر و ۲۰۰ ماده) به‌طور تصادفی از نسل اول به‌عنوان والدین نسل دوم انتخاب شدند، به‌طوری که تعداد افراد در نسل دوم جمعیت اخیر به ۱۰۰۰ فرد رسید. این روند برای نسل سوم هم ادامه یافت تا افراد در این نسل هم به ۱۰۰۰ فرد برسد. نسل دوم جمعیت اخیر به‌عنوان افراد مرجع در نظر گرفته شد، در نتیجه، این افراد دارای اطلاعات فنوتیپی و ژنوتیپی نشانگری بودند. افراد نسل سوم، جمعیت تایید را تشکیل دادند که فقط اطلاعات ژنوتیپی برای این افراد شبیه‌سازی شد و GEBV برای این افراد پیش‌بینی شد. در نسل ۱۰۰۰، مقدار آماره $r^2$ به‌عنوان معیاری از سطح LD برای مکان-های مجاور در ژنوم با فرمول زیر برآورد شد (Hill and Rabertson, 1968):

$$r^2 = \frac{D^2}{\text{freq}(A_1) \times \text{freq}(A_2) \times \text{freq}(B_1) \times \text{freq}(B_2)}$$
$$D = \text{freq}(A_1B_1) \times \text{freq}(A_2B_2) - \text{freq}(A_1B_2) \times \text{freq}(A_2B_1)$$

که در آن، $r^2$ مجذور ضریب همبستگی بین دو جایگاه A و B است، $freq(A_1)$، $freq(A_2)$، $freq(B_1)$ و $freq(B_2)$، به‌ترتیب فراوانی آلل‌های $A_1$، $A_2$، $B_1$ و $B_2$ در جمعیت است. همچنین، $freq(A_1B_1)$، $freq(A_1B_2)$، $freq(A_2B_1)$ و $freq(A_2B_2)$ به‌ترتیب فراوانی هاپلوتایپ‌های $A_1B_1$، $A_1B_2$، $A_2B_1$ و $A_2B_2$ هستند. D نیز انحراف ژنوتیپ‌های والدین از ژنوتیپ‌های نوترکیب است. در این پژوهش، ژنومی شامل شش کروموزوم با طول ۱۰۰ سانتی مورگان با واریانس فنوتیپی ($\sigma_p^2$) و وراثت پذیری خاص ($h^2$) به‌ترتیب برابر با ۱ و ۰/۴ شبیه‌سازی شد. روی هر کروموزوم، ۱۰۰۰ SNP با فواصل یکسان و ۱۰۰ QTL به‌طور تصادفی تعبیه شدند. آثار افزایشی QTLها از توزیع گاما با پارامتر شکل برابر با ۰/۴ و پارامتر مقیاس برابر با ۱/۶۶، نمونه‌گیری شد (Meuwissen *et al.,* 2001). همچنین، درجه غالبیت از توزیع نرمال با میانگین ۰/۵ و واریانس ۱ نمونه‌گیری شد (Esfandiari and Sorensen, 2017). واریانس انحراف غالبیت ($\sigma_d^2$) در سناریوهای مختلف برابر با ۰/۱۰، ۰/۱۵، ۰/۲۰، ۰/۲۵، ۰/۳۰ و ۰/۳۵ در نظر گرفته شد. نرخ جهش برای جایگاه‌های نشانگر و QTL، $2.5\times10^{-5}$ درنظر گرفته شد. انتخاب شش کروموزوم و سطوح مختلف واریانس غالبیت برای شبیه‌سازی یک معماری ژنتیکی جامع و در عین حال واقعی برای صفات بود که این امکان را می‌داد تا تاثیر روش‌های آماری مختلف را بر صحت پیش‌بینی ژنومی در یک حالت کلی ارزیابی کنیم. هدف این بود که شبیه‌سازی را برای انواع گونه‌ها و صفات بدون تمرکز روی سازواره خاص، قابل تفسیر کنند. شبیه‌سازی شش کروموزوم به‌عنوان یک مدل ساده و کلی انتخاب شد که می‌تواند گونه‌های مختلف دام یا گیاه را در مطالعات ژنتیکی نشان دهد. این رویکرد، در پژوهش‌های دیگر نیز مورد استفاده قرار گرفته است (Ghafouri-Kesbi *et al.,* 2016; Sahebalam *et al.,* 2022). فنوتیپ‌ها با دو مدل کنش ژنی مختلف شبیه‌سازی شدند: *مدل صرفاً افزایشی*: اثر متوسط آلل $A_1$ به‌صورت $q[a + d(q - p)]$ و اثر متوسط آلل $A_2$ به‌صورت $-p[a + d(q - p)]$ محاسبه می‌شود. اثر متوسط جایگزینی ژن از انحراف ارزش‌های متوسط دو آلل حاصل می‌شود که با نماد $\alpha$ نمایش داده می‌شود و به‌صورت زیر به‌دست آمد:

$$\alpha = \alpha_1 - \alpha_2 = q[a + d(q - p)] - (-p[a + d(q - p)]) = [a + d(q - p)](q + p) = a + d(q - p)$$

که $a$ و $d$ به‌ترتیب آثار افزایشی و غالبیت هستند و $p$ فراوانی آلل $A_1$ و $q = 1 - p$ است. به‌منظور ایجاد صفت صرفاً افزایشی، در ابتدا باید آثار افزایشی و انحراف غالبیت شبیه‌سازی شوند تا اثر متوسط جایگزینی هر QTL به‌دست آید. ارزش‌های ژنتیکی افزایشی از مجموع حاصلضرب کد‌های ژنوتیپی در آثار متوسط جایگزینی QTL به‌دست می‌آید. ارزش فنوتیپی هر فرد i ($y_i$) با اضافه کردن یک اثر باقیمانده دارای توزیع نرمال $e_i \sim N(0, \sigma_e^2)$ به‌مجموع QTLها (ارزش‌های ژنتیکی) به‌صورت زیر به‌دست می‌آید:

$$y_i = \sum_{k}^{nQTL} X_{ik}\alpha_k + e_i$$

که در آن، $X_{ik}$ ($i$=1,…,n و $k$=1,…,n) یک عنصر از ماتریس طرح برای آثار متوسط جایگزینی QTL ($\alpha_k$) و $e_i$ اثر تصادفی باقیمانده است که $\sigma_e^2$ واریانس باقیمانده است. ژنوتیپ‌ها برای به دام انداختن آثار افزایشی به‌صورت $A_2A_2$ برای ۲-۲p، $A_1A_2$ برای ۱-۲p و $A_1A_1$ برای ۰-۲p کدگذاری شدند. ارزش‌های اصلاحی برای ژنوتیپ $A_1A_1$ به‌صورت $2q\alpha$، برای ژنوتیپ $A_1A_2$ به‌صورت $(q - p)\alpha$ و برای ژنوتیپ $A_2A_2$ به‌صورت $-2p\alpha$ است.

*مدل افزایشی + انحراف غالبیت*: غالبیت زمانی رخ می‌دهد که اثر آلل‌ها در یک جایگاه ژنی به‌گونه‌ای باشد که ارزش ژنوتیپ هتروزیگوت از میانگین ارزش ژنوتیپ هموزیگوت‌ها، انحراف داشته باشد. انحراف غالبیت برای یک QTL خاص، به‌صورت تفاوت بین ارزش متوسط ژنوتیپ $A_1A_2$ و میانگین ژنوتیپ‌های $A_1A_1$ و $A_2A_2$ بیان می‌شود. انحراف غالبیت برای یک جایگاه ژنی، از تفاوت بین ارزش ژنوتیپی کل و ارزش اصلاحی به‌دست می‌آید و برابر با $-2q^2d$ برای ژنوتیپ $A_1A_1$، $2pqd$ برای ژنوتیپ $A_1A_2$ و $-2p^2d$ برای ژنوتیپ $A_2A_2$ است (Falconer and Mackay, 1996). برای شبیه‌سازی آثار غالبیت، ابتدا درجه غالبیت $h_i$ از توزیع نرمال $h_i \sim N(0.5,1)$ نمونه‌گیری شد (Esfandiari and Sorensen, 2017). سپس، آثار غالبیت به‌صورت $d_k = h_k.|a_k|$ در نظر گرفته شد که $|a_k|$ ارزش مطلق اثر افزایشی است. فنوتیپ‌های صفت با اضافه شدن اثر

باقیمانده نرمال استاندارد به‌مجموع ارزش اصلاحی واقعی و انحراف غالبیت هر فرد ایجاد شدند:

$$y_i = \sum_{k}^{nQTL} (X_{ik}a_k + D_{ik}d_k) + e_i$$

که $D_{ik}$ (i=1,…,n و k=1,…,nQTL) یک عنصر از ماتریس طرح برای آثار اتحراف غالبیت ($d_k$) است که به‌صورت $-2q^2$، $2pq$ و $-2p^2$ به‌ترتیب برای ژنوتیپ‌های $A_1A_1$، $A_1A_2$ و $A_2A_2$ کدگذاری می‌شود.

*بهترین پیش‌بینی کننده نااریب خطی (GBLUP)*: در این روش، از معادلات مدل مختلط BLUP معمول استفاده شد، فقط با این تفاوت که به‌جای استفاده از معکوس ماتریس شجره ($\mathbf{A^{-1}}$) از معکوس ماتریس روابط ژنومی ($\mathbf{G^{-1}}$) استفاده می‌شود (Habier *et al.*, 2007; Hayes *et al.*, 2009). در این روش، آثار نشانگری از یک توزیع نرمال نمونه‌گیری می‌شوند.

مدل خطی استفاده شده برای مدل کنش ژنی صرفاً افزایشی به‌صورت زیر است:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Tu} + \boldsymbol{e}$$

در این رابطه، **y** بردار مشاهدات فنوتیپی، **b** بردار آثار ثابت (میانگین کل) و **u** بردار آثار ژنتیکی افزایشی که از توزیع نرمال $N(0,\mathbf{G}\sigma_a^2)$ پیروی می‌کند، **G** ماتریس روابط خویشاوندی افزایشی ژنومی و $\sigma_a^2$ واریانس آثار افزایشی است. **X** و **T** ماتریس های طرح به‌ترتیب برای **b** و **u** است.

مدل خطی استفاده شده برای مدل کنش ژنی افزایشی-انحراف غالبیت به‌صورت زیر بود:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Tu} + \boldsymbol{Td} + \boldsymbol{e}$$

در این رابطه، **d** بردار آثار ژنتیکی غالبیت که از یک توزیع نرمال به‌صورت $d\sim N(0,\mathbf{D}\sigma_d^2)$ تبعیت می‌کند، **D** ماتریس روابط غالبیت ژنومی و $\sigma_d^2$ واریانس ژنتیکی غالبیت است. ماتریس روابط خویشاوندی ژنومی افزایشی (**G**) به‌صورت زیر برآورد شد (VanRaden, 2008):

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{2\sum_{j=1}^{m} p_j(1-p_j)}$$

در این رابطه، **Z** ماتریس مرکزی‌شده به‌وسیله **M-P** است، که **M** ماتریس ژنوتیپی است که به‌صورت ۰، ۱ و ۲ مطابق با تعداد آلل‌های جایگزین کدگذاری می‌شود. **P** ماتریس فراوانی آللی است که به‌صورت $2p_j$ است، $p_j$ jامین فراوانی آلل جایگزین و $2\sum_{j=1}^{m} p_j(1-p_j)$ مجموع واریانس افزایشی سراسر جایگاه‌های ژنی است. ماتریس روابط خویشاوندی ژنومی انحراف غالبیت (D) به‌صورت زیر برآورد شد (Vitezica *et al.*, 2013):

$$\boldsymbol{D} = \frac{\boldsymbol{WW'}}{\sum_{j=1}^{m}\left(2p_j(1-p_j)\right)^2}$$

در این رابطه، **W** دارای ابعادی برابر با ماتریس **Z** است که عناصر آن برابر با $-2q_j^2$، $2p_jq_j$ و $-2p_j^2$ به‌ترتیب برای ژنوتیپ‌های $A_1A_1$ و $A_1A_2$، $A_2A_2$ است. $\sum_{j=1}^{m}\left(2p_j(1-p_j)\right)^2$ مجموع واریانس انحراف غالبیت سراسر جایگاه‌ها است.

*ماشین بردار پشتیبان (Support Vectoe Machine)*: روش ناپارامتری ماشین بردار پشتیبان (SVM) (Cottex and Vapnik, 1995)، یک روش یادگیری تحت نظارت است که در ابتدا به‌عنوان یک طبقه‌بندی کننده توسعه پیدا کرد. یک مجموعه داده آموزشی برای ایجاد یک طبقه‌بندی کننده حداکثر حاشیه بکار برده شد که بیشترین تفاوت ممکن را در بین دو کلاس از مشاهدات ایجاد می‌کند. در مورد تفکیک‌پذیری خطی، اگر مشاهدات $(x_i) \in R^P$ باشد، آنگاه تفکیک‌کننده، یک ابر-صفحه (hyper-plane) در $R^{P-1}$ است. از آنجا که برازش مدل رگرسیونی اساساً شامل یافتن تجسم بهینه از مشاهدات در یک ابر-صفحه با ابعاد کمتر است، می توان از این ایده برای برآورد تابع رگرسیون ناشناخته با محدودیت استفاده کرد. یکی از ویژگی‌های خوب رگرسیون SVM در کاربردهای اصلاح دام و نباتات این است که روابط بین ژنوتیپ‌های نشانگر و فنوتیپ‌ها را می‌توان با یک تابع نقشه (mapping function) خطی یا غیرخطی، مدل‌سازی کرد، که نمونه‌ها را از فضای پیش‌بینی کننده (predictor space) به یک فضای ویژگی چند-بعدی می‌برد. با فرض یک نمونه آموزشی $S=\{(x_i, y_i), x_i \in R^n, y_i \in R, i=1\ldots n\}$، که $x_i$ یک بردار p-بعدی که شامل ارزش-های ژنوتیپی برای p نشانگر برای فرد iام، است، $y_i$ ارزش فنوتیپی برای فرد iام است. مدلی که رابطه بین فنوتیپ و ژنوتیپ یک فرد را توصیف می‌کند می‌تواند به‌صورت زیر نوشته شود:

$$f(x) = b + \boldsymbol{w}x$$

که b یک فراسنجه ثابت و **w** بردار وزن‌های ناشناخته یا ضرایب رگرسیونی است. ثابت b بیانگر حداکثر خطایی است که در هنگام برآورد وزن **w** رخ می‌دهد. یادگیری $f(x)$ با به حداقل رساندن عبارت زیر به‌دست می‌آید:

$$\lambda \sum_{i=1}^{n} L\left(y_i - f(x_i)\right) + \frac{1}{2}\|w\|^2$$

که $L(.)$ تابع زیان (loss function) است که کیفیت برآورد را اندازه می‌گیرد. فراسنجه تنظیم‌کننده $\lambda$ بین کم‌تراکمی (sparsity) و پیچیدگی مدل، موازنه ایجاد می‌کند. افزایش $\lambda$ منجر به جریمه بیشتر روی خطا می‌شود. نرم $\|w\|$ از بردار **w** رابطه معکوسی با پیچیدگی مدل دارد، با انتخاب **w** برای به حداقل رساندن $\|w\|$، می‌توان پیچیدگی مدل را کاهش داد. توابع زیان بسیاری برای رگرسیون SVM استفاده شده است. برخی از این توابع زیان شامل زیان مربع (squared loss)، زیان مطلق (absolute loss) و زیان غیرحساس-ε (ε-insensitive loss) هستند. تابع زیان مربع دارای فرم $L(y-f(x)) = (y-f(x))^2$ است. این تابع میزان زیان را به‌صورت درجه دوم با اندازه خطا مقیاس می‌کند. استفاده از این تابع زیان نشان می‌دهد که داده‌های پرت نیز به‌صورت درجه دوم وزن‌دهی می‌شوند، که نیاز است کاربر قبل از تجزیه رگرسیون با داده‌های پرت مقابله کند. تابع زیان مطلق دارای فرم $L(y-f(x)) = |(y-f(x))|$ است. تابع زیان مطلق، میزان زیان را به‌صورت خطی با میزان خطا مقیاس می‌کند که مشکل استفاده از مجموعه داده‌هایی که داده‌های پرت دارند را از بین می‌برد.

تابع زیان ε-غیرحساس دارای فرم زیر است:

$$L(y-f(x)) = \begin{cases} 0 & if\ |y-f(x)| < \varepsilon \\ |y-f(x)| - \varepsilon & otherwise \end{cases}$$

که ε، تعداد بردارهای پشتیبان استفاده شده در تابع رگرسیون را تعیین می‌کند. بر اساس تعریف (Vapnik, 1995)، یک بردار پشتیبان، بردار $x_i$ است که معادله $y_i(wx_i + b) = 1$ را ایجاد کند. افزایش ε نشان می‌دهد که بردارهای پشتیبان کمتری در برازش استفاده می‌شوند. تابع زیان ε-غیرحساس، خطاها را در رگرسیون که اندازه آنها کمتر از ε است، نادیده می‌گیرد. زمانی‌که خطا بزرگتر از ε است، تابع زیان به‌صورت $|y-f(x)| - \varepsilon$ می‌شود. تابع زیان ε-غیرحساس نیازمند یک نمایش توانمندتری برای محاسبه مقدار خطا در داده‌ها است. بدین منظور می‌توان یک هزینه بیشتر یا یک عدم اطمینان بیشتر را با معرفی متغیرهای کمکی نامنفی $\xi$ محدود شده، به تابع زیان اضافه کرد (Long, 2011): $\xi_{1i} \geq y_i - f(x_i) -$ ε، که i=1…n است که n تعداد مشاهدات آموزشی است، $\xi_{2i} \geq f(x_i) - y_i -$ ε، که i=1…n است، اکنون با در نظر گرفتن این دو فرض، می‌توان تابع هدف را به حداقل رساند:

$$\lambda \sum_{i=1}^{n} \left(\xi_{1i} + \xi_{2i}\right) + \frac{1}{2}\|w\|^2$$

راه‌حل این مشکل به حداقل رساندن محدود شده به فرم زیر است (Nocedal and Wright, 1999):

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_1 x_i x + b$$

این فرمول وابسته به داده‌های آموزشی از طریق ضرب داخلی (inner product) $\langle x_i, x_j \rangle$ است، که یک تابع خطی از مشاهدات است. برای استفاده از فضاهای ویژگی ابعاد بالاتر، می‌توان داده‌ها را از طریق توابع غیرخطی معرفی کرد. برای مثال، می‌توان ضرب داخلی داده‌ها را با یک تابع کرنل جایگزین کرد:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

به‌طور کلی، یک کرنل را می‌توان به‌صورت $k(x,z)=\phi(x)^T\phi(z)$ بیان کرد که x و z دو بردار در فضای اصلی و $\phi(x)$ و $\phi(z)$ بردارهایی در فضای ویژگی هستند. *روش‌های مورد ارزیابی:* تابع کرنل خطی در مدل ماشین بردار پشتیبان به‌صورت زیر است (SVM-linear): $K(x,z)=xz$. تابع کرنل پایه شعاعی گاوسی در مدل ماشین بردار پشتیبان به‌صورت زیر است (SVM-radial): $K(x,z)=\exp(-\sigma\|x-z\|^2)$ که σ فراسنجه پهنای باند (بندویدج) است. σ یک فراسنجه خاص هسته است که اجازه می‌دهد تا یک تابع خطی در یک فضای ویژگی بی‌نهایت بزرگ پیدا شود. تابع کرنل چندجمله‌ای در مدل ماشین بردار پشتیبان به‌صورت زیر است (SVM-polynomial): $K(x,z)=(\sigma xz+\vartheta)^d$. در این رابطه، $\vartheta$ عرض از مبداء و $d$ درجه چندجمله‌ای است. تابع کرنل حلقوی در مدل ماشین بردار پشتیبان به‌صورت زیر است (SVM-sigmoid): $K(x,z)=\tanh(\sigma xz+\vartheta)$. با توجه به انتخاب تابع کرنل، مدل رگرسیون بردار پشتیبان غیرخطی حاصل، ترکیبی خطی از تابع کرنل به‌صورت زیر است:

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{a}_i k(x_i, x_j) + \hat{b}$$

*اجرای روش‌های آماری:* روش GBLUP با استفاده از بسته نرم افزاری BGLR (Perez and de los Campos, 2014) و روش ماشین بردار پشتیبان بر اساس کرنل‌های خطی

(SVM-lin)، شعاعی گاوسی (SVM-rad)، چندجمله‌ای (SVM-pol) و حلقوی (SVM-sig)، بااستفاده از بسته نرم افزاری e1071 (Meyer *et al.*, 2019) در محیط نرم افزاری R نسخه 4.3.2 اجرا شد. به دلیل استفاده از مدل تصادفی و پایداری بیشتر برآوردها، هر سناریو برای ۱۰۰ بار تکرار شد. میانگین و انحراف معیار نتایج برای مقایسه‌های لازم ارائه شد. کنترل کیفیت داده‌های ژنوتیپی با استفاده از فراوانی آلل کمیاب (MAF) انجام شد. در این مورد، برای هر تکرار، نشانگرهایی با MAF کمتر از ۰/۰۵ قبل از برآورد آثار نشانگر از ماتریس ژنوتیپی حذف شدند.

*صحت پیش‌بینی:* صحت پیش‌بینی به‌عنوان ضریب همبستگی پیرسون بین ارزش ژنتیکی واقعی (TGV) یا ارزش افزایشی واقعی (TBV) و ارزش ژنتیکی پیش‌بینی شده ژنومی (GEGV) یا ارزش افزایشی پیش‌بینی شده ژنومی (GEBV) تعریف شد، یعنی به‌ترتیب r(TGV,GEGV) و r(TBV,GEBV). زمانی‌که پیش‌بینی‌های مدل فقط آثار ژنتیکی افزایشی بود، از تخمین‌های GEBV استفاده شد، در حالی که برای مدل‌های افزایشی + انحراف غالبیت به‌طور جداگانه از برآوردهای GEBV و GEGV برای محاسبه صحت پیش-بینی استفاده شد. TBVها و TGVها در دو مدل کنش ژنی صرفاً افزایشی و افزایشی + انحراف غالبیت، شبیه‌سازی شدند، درنتیجه کاملاً مشخص بودند.

*معنی‌داری عملی:* برای بررسی میزان فاصله (شباهت یا اختلاف) صحت‌های پیش‌بینی GEBV و GEGV رویکردهای آماری مورد مطالعه در طول ۱۰۰ تکرار از معنی‌داری عملی بر اساس اندازه اثر d کوهن استفاده شد. آماره d کوهن (Cohen, 1988) می‌تواند برای توصیف تفاوت میانگین استاندارد شده یک اثر یا مقایسه آثار در مطالعات مختلف استفاده شود و به روش‌های مختلف اندازه‌گیری متغیر وابسته بستگی ندارد. آماره d که تفاوت استاندارد شده بین دو گروه مشاهدات مستقل برای نمونه است، به-صورت زیر محاسبه می‌شود:

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{S_P}$$

که در این رابطه، صورت کسر بیانگر تفاوت بین میانگین دو گروه مشاهدات و مخرج آن، انحراف معیار جمع شده است. بازه عددی آن از ۰ تا بی‌نهایت، متغیر است. d=0.2، d=0.5 و d=0.8 به‌ترتیب نشان‌دهنده اندازه اثر کوچک، متوسط و بزرگ است. اندازه اثر کوچک به‌معنای یک رابطه یا تفاوت ضعیف است، در حالی‌که اندازه اثر بزرگ نشان‌دهنده یک رابطه یا تفاوت قوی است که حاکی از اهمیت عملی است.

## نتایج

*کنترل کیفیت و عدم تعادل پیوستگی (LD):* بر اساس داده شبیه‌سازی شده، افراد جمعیت مرجع و تایید دارای ماتریس ژنوتیپی با ابعاد ۶۰۰۰×۱۰۰۰ بودند. از ۶۰۰۰ نشانگر بعد از کنترل کیفیت بر اساس فراوانی آلل کمیاب (MAF<0.05)، ۳۸۶۴ نشانگر به‌طور میانگین در ۱۰۰ تکرار به‌عنوان پیش‌بینی‌کننده در مدل‌های آماری مختلف روی مشاهدات برازش داده شد، در نتیجه، ابعاد این ماتریس به ۱۰۰۰×۳۸۶۴ کاهش یافت. آماره $r^2$ که به‌عنوان معیاری برای ارائه سطح LD برای موقعیت‌های مجاور در ژنوم در نظر گرفته می‌شود، حدود ۰/۲۲، به‌طور میانگین در ۱۰۰ تکرار، برآورد شد.

*صحت پیش‌بینی GEBV و GEGV:* صحت پیش‌بینی GEBV و GEGV روش‌های GBLUP و ماشین بردار پشتیبان بر اساس کرنل‌های مختلف خطی (SVM-lin)، شعاعی (SVM-rad)، چندجمله ای (SVM-pol) و حلقوی (SVM-sig) با در نظر گرفتن واریانس غالبیت مختلف در جدول ۱ و شکل‌های ۱ و ۲ ارائه شده است. همچنین، خروجی آزمون d کوهن برای بررسی فاصله بین صحت پیش‌بینی GEBV یا GEGV رویکردهای مختلف مورد-مطالعه براساس ارزش‌های مختلف واریانس غالبیت، به-ترتیب در جداول ۲ و ۳ ارائه شده است. بر اساس نتایج به‌دست آمده، رویکرد GBLUP توانست صحت پیش‌بینی GEBV و GEGV بالاتری را بر اساس ارزش‌های مختلف واریانس غالبیت گزارش دهد، به‌طوری که با افزایش این واریانس، اختلاف بین عملکرد این روش با رویکردهای مختلف SVM کاهش یافت. در مدل صرفاً افزایشی ($\sigma_d^2 = 0$)، رویکرد SVM-rad بالاترین صحت پیش‌بینی GEBV و GEGV را با اختلاف نسبت به سایر رویکردهای SVM نشان داد. همچنین، پایین‌ترین عملکرد در SVM-lin مشاهده شد. با در نظر گرفتن حداقل واریانس غالبیت ($\sigma_d^2$=0.1)، صحت پیش‌بینی GEBV در روش‌های GBLUP و SVM-rad کاهش یافت. همچنین، این سطح از واریانس غالبیت، منجر به افزایش چشمگیر این صحت به‌ترتیب در SVM-lin و SVM-pol شد. بر اساس صحت

پیش‌بینی GEGV، افت چشمگیری در تمام روش‌های مورد مطالعه نسبت به مدل صرفاً افزایشی مشاهده شد. بالاترین صحت پیش‌بینی GEBV و GEGV در مدل SVM-rad مشاهده شد. همچنین، پایین‌ترین صحت GEBV و GEGV به‌ترتیب در رویکردهای SVM-pol و SVM-lin گزارش شد. با افزایش سهم غالبیت در شکل‌گیری فنوتیپ صفت، صحت پیش‌بینی GEBV در مدل افزایشی-انحراف غالبیت، در تمام رویکردها افزایش یافت. زمانی‌که واریانس غالبیت نقش بزرگی را در واریانس فنوتیپی ایفا کرد ($\sigma^2_d$=0.30 و $\sigma^2_d$=0.35)، تمام رویکردها، صحت GEBV بالاتری را نسبت به مدل صرفاً افزایشی نشان دادند، به‌استثنای رویکرد SVM-rad که تقریباً صحت یکسانی را ثبت کرد. رویکرد SVM-lin که در مدل صرفاً افزایشی پایین‌ترین صحت را نشان داده بود، در این سطح از واریانس غالبیت توانست بالاترین عملکرد را در بین رویکردهای مختلف SVM نشان دهد. علاوه بر این، ارزش‌های مختلف واریانس غالبیت نتوانست باعث عملکرد بهتر رویکردهای مورد مطالعه، بر اساس GEGV شود، به‌طوری که، با در نظر گرفتن حداکثر واریانس غالبیت، بالاترین و پایین‌ترین عملکرد در بین رویکردهای خانواده SVM، به‌ترتیب مربوط به رویکردهای SVM-rad و SVM-pol بود. میزان فاصله (یا اختلاف) صحت پیش‌بینی GEBV و GEGV بین رویکردهای GBLUP و SVM-rad در مدل صرفاً افزایشی، حداقل (d=0.218) و در مدل افزایشی-انحراف غالبیت با حداکثر واریانس غالبیت ($\sigma^2_d$=0.35)، حداکثر (به‌ترتیب d=0.492 و d=0.404) بود، که دلیل این امر این است که در مدل افزایشی-انحراف غالبیت با افزایش واریانس غالبیت، تغییر صحت در روش GBLUP اندکی بیشتر از روش SVM-rad بود. رویکرد SVM-lin توانست در مدل افزایشی-انحراف غالبیت نسبت به مدل صرفاً افزایشی، فاصله صحت پیش‌بینی GEBV خود را نسبت به روش‌های GBLUP و SVM-rad به‌شدت کاهش دهد، به‌طوری که با در نظر گرفتن حداکثر واریانس غالبیت، این فاصله به‌ترتیب برابر با d=0.309 و d=0.189 گزارش شد. به‌طورکلی، بیشترین فاصله صحت پیش‌بینی بین رویکردهای مورد مطالعه (بجز بین رویکردهای GBLUP و SVM-rad و همچنین، بین SVM-lin و SVM-pol) در مدل صرفاً افزایشی مشاهده شد، به‌طوری که در مدل افزایشی-انحراف غالبیت با افزایش واریانس غالبیت، این فاصله تنزل پیدا کرد. در بین رویکردهای خانواده SVM، کمترین فاصله صحت پیش‌بینی GEBV به‌ترتیب بین رویکردهای SVM-lin و SVM-sig (d=0.020) و بین SVM-lin و SVM-rad (d=0.030) به‌ترتیب در واریانس غالبیت برابر با ۰/۱۵ و ۰/۳۰ ثبت شد. علاوه بر این، کمترین فاصله صحت پیش‌بینی GEGV هم بین SVM-lin و SVM-pol (d=0.010) در واریانس غالبیت برابر با ۰/۱۵ گزارش شد. همچنین، بیشترین فاصله بین رویکردهای SVM-lin و SVM-rad (d=2.337) و همچنین، بین SVM-rad و SVM-pol (d=1.939) در مدل صرفاً افزایشی مشاهده شد.

نمودار جعبه‌ای، صحت پیش‌بینی GEBV (شکل ۱) و GEGV (شکل ۲) را برای روش‌های مختلف با استفاده از ماشین بردار پشتیبان با توابع هسته‌ای متفاوت و مقایسه آن‌ها با روش رایج GBLUP در سطوح مختلف واریانس غالبیت را نشان می‌دهد. به‌طورکلی، روش‌های SVM نسبت به GBLUP دارای صحت پایین‌تری بودند و صحت‌های پیش‌بینی GEBV و GEGV همه روش‌ها با افزایش واریانس غالبیت به‌ترتیب روند تقریباً افزایشی و کاهشی نشان دادند. بر اساس نمودار جعبه‌ای شکل ۱، تابع SVM-lin بهترین عملکرد را بر اساس صحت GEBV در بیشترین سطح واریانس غالبیت داشت، درحالی‌که به‌ترتیب توابع SVM-rad، SVM-sig و SVM-pol، کمترین اختلاف را با این روش نشان دادند. بر اساس نمودار جعبه‌ای شکل ۲، روش SVM-rad بهترین عملکرد را در عدم وجود واریانس غالبیت نشان داد، و به‌ترتیب روش‌های SVM-sig، SVM-pol و SVM-lin، کمترین اختلاف را با این روش داشتند. نقاط درج شده در نمودار نشان‌دهنده پراکندگی داده‌ها و وجود داده‌های پرت هستند که نشان می‌دهد تغییرات زیادی در صحت پیش‌بینی وجود دارد. این تحلیل نشان می‌دهد که انتخاب تابع هسته‌ای مناسب در ماشین‌های بردار پشتیبان برای بهبود صحت پیش‌بینی GEBV و GEGV بسیار مهم است.

## بحث

ارزش متوسط عدم تعادل پیوستگی ($r^2$)، مجذور ضریب همبستگی بین دو مکان مجاور است و نشان‌دهنده نسبت واریانس QTL است که به‌وسیله نشانگرها توجیه می‌شود (Goddard and Hayes, 2007). توصیه شده است که سطح عدم تعادل پیوستگی مورد نیاز برای یک پروژه اصلاح

ژنومی موفق حدود ۲۰٪ باشد (Meuwissen *et al.*, 2001). بر اساس نتایج به‌دست آمده، رویکردهای SVM-rad و SVM-sig، بالاترین صحت پیش‌بینی را نشان دادند، که با پژوهش قبلی مطابقت داشت (Kasnavi *et al.*, 2018). محققین بر اساس صفت گسسته شبیه‌سازی شده، نشان دادند که پیش‌بینی‌کننده‌های SVM مبتنی بر شعاعی و سیگموئید، عملکرد بهتری از کرنل‌های چندجمله‌ای و خطی داشتند. همچنین، توصیه کردند که به‌دلیل ارائه پیش‌بینی‌هایی با حداقل صحت به‌وسیله کرنل‌های خطی و چندجمله‌ای، از این کرنل‌ها برای پیش‌بینی گسترده ژنوم صفات گسسته استفاده نشود.

جدول ۱- صحت پیش‌بینی GEBV و GEGV روش GBLUP و رویکردهای مختلف ماشین بردار پشتیبان بر اساس توابع کرنل خطی (SVM-lin)، شعاعی (SVM-rad)، چندجمله ای (SVM-pol) و حلقوی (SVM-sig) در سطوح مختلف واریانس غالبیت ($\sigma_d^2$)

Table 1. GEBV and GEGV prediction accuracy of GBLUP method and different support vector machine approaches based on linear (SVM-lin), radial (SVM-rad), polynomial (SVM-pol), and cyclic (SVM-sig) kernel functions at different levels of dominance variance ($\sigma_d^2$)

| Prediction models | Accuracy | $\sigma_d^2$=0.00 | $\sigma_d^2$=0.10 | $\sigma_d^2$=0.15 | $\sigma_d^2$=0.20 | $\sigma_d^2$=0.25 | $\sigma_d^2$=0.30 | $\sigma_d^2$=0.35 |
|---|---|---|---|---|---|---|---|---|
| GBLUP | r(TBV,GEBV) | 0.554 (0.044) | 0.537 (0.047) | 0.539 (0.062) | 0.544 (0.048) | 0.552 (0.050) | 0.559 (0.048) | 0.569 (0.053) |
| | r(TGV,GEGV) | 0.554 (0.044) | 0.460 (0.059) | 0.440 (0.076) | 0.429 (0.064) | 0.439 (0.056) | 0.436 (0.054) | 0.444 (0.063) |
| SVM-lin | r(TBV,GEBV) | 0.431 (0.051) | 0.499 (0.051) | 0.507 (0.070) | 0.515 (0.053) | 0.530 (0.051) | 0.541 (0.049) | 0.554 (0.047) |
| | r(TGV,GEGV) | 0.431 (0.052) | 0.365 (0.054) | 0.362 (0.071) | 0.360 (0.058) | 0.376 (0.048) | 0.380 (0.049) | 0.387 (0.057) |
| SVM-rad | r(TBV,GEBV) | 0.546 (0.048) | 0.526 (0.047) | 0.527 (0.062) | 0.532 (0.051) | 0.540 (0.050) | 0.542 (0.047) | 0.545 (0.048) |
| | r(TGV,GEGV) | 0.546 (0.048) | 0.441 (0.056) | 0.423 (0.067) | 0.414 (0.061) | 0.421 (0.052) | 0.417 (0.052) | 0.419 (0.058) |
| SVM-pol | r(TBV,GEBV) | 0.451 (0.048) | 0.460 (0.050) | 0.463 (0.069) | 0.471 (0.054) | 0.473 (0.054) | 0.485 (0.045) | 0.490 (0.053) |
| | r(TGV,GEGV) | 0.451 (0.048) | 0.375 (0.056) | 0.362 (0.075) | 0.347 (0.067) | 0.346 (0.069) | 0.359 (0.058) | 0.355 (0.070) |
| SVM-sig | r(TBV,GEBV) | 0.503 (0.050) | 0.504 (0.052) | 0.508 (0.065) | 0.521 (0.051) | 0.527 (0.051) | 0.533 (0.048) | 0.542 (0.048) |
| | r(TGV,GEGV) | 0.503 (0.050) | 0.394 (0.058) | 0.384 (0.071) | 0.382 (0.060) | 0.391 (0.051) | 0.392 (0.051) | 0.396 (0.058) |

جدول ۲- خروجی اندازه اثر d کوهن به‌منظور بررسی فاصله بین صحت پیش‌بینی GEBV روش‌های مختلف GBLUP، SVM-lin، SVM-rad، SVM-pol و SVM-sig در سراسر ۱۰۰ تکرار

Table 2. Output of the Cohen's d effect size to examine the distance between the prediction accuracy of GEBV of GBLUP, SVM-lin, SVM-rad, SVM-pol, and SVM-sig across 100 iterations

| Distance | Dominance variance | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\sigma_d^2$=0.00 | $\sigma_d^2$=0.10 | $\sigma_d^2$=0.15 | $\sigma_d^2$=0.20 | $\sigma_d^2$=0.25 | $\sigma_d^2$=0.30 | $\sigma_d^2$=0.35 |
| GBLUP: SVM-lin | 2.608 | 0.774 | 0.475 | 0.558 | 0.471 | 0.366 | 0.309 |
| GBLUP: SVM-rad | 0.218 | 0.235 | 0.185 | 0.229 | 0.284 | 0.344 | 0.492 |
| GBLUP: SVM-pol | 2.201 | 1.581 | 1.153 | 1.422 | 1.567 | 1.575 | 1.548 |
| GBLUP: SVM-sig | 1.118 | 0.675 | 0.471 | 0.454 | 0.543 | 0.532 | 0.553 |
| SVM-lin: SVM-rad | 2.337 | 0.551 | 0.302 | 0.325 | 0.192 | 0.030 | 0.189 |
| SVM-lin: SVM-pol | 0.400 | 0.771 | 0.638 | 0.840 | 1.093 | 1.190 | 1.278 |
| SVM-lin: SVM-sig | 1.431 | 0.083 | 0.020 | 0.107 | 0.690 | 0.161 | 0.251 |
| SVM-rad: SVM-pol | 1.939 | 1.358 | 0.980 | 1.173 | 1.297 | 1.245 | 1.090 |
| SVM-rad: SVM-sig | 0.882 | 0.455 | 0.292 | 0.220 | 0.263 | 0.193 | 0.061 |
| SVM-pol: SVM-sig | 1.036 | 0.842 | 0.679 | 0.955 | 1.030 | 1.036 | 1.033 |

جدول ۳- خروجی اندازه اثر d کوهن به‌منظور بررسی فاصله بین صحت پیش‌بینی GEGV روش‌های مختلف GBLUP، SVM-lin، SVM-rad، SVM-pol و SVM-sig در سراسر ۱۰۰ تکرار

Table 3. Output of the Cohen's d effect size to examine the distance between the prediction accuracy of GEGV of GBLUP, SVM-lin, SVM-rad, SVM-pol, and SVM-sig across 100 iterations

| Distance | Dominance variance | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\sigma_d^2$=0.00 | $\sigma_d^2$=0.10 | $\sigma_d^2$=0.15 | $\sigma_d^2$=0.20 | $\sigma_d^2$=0.25 | $\sigma_d^2$=0.30 | $\sigma_d^2$=0.35 |
| GBLUP: SVM-lin | 2.608 | 1.664 | 1.053 | 1.128 | 1.217 | 1.094 | 0.948 |
| GBLUP: SVM-rad | 0.218 | 0.328 | 0.239 | 0.234 | 0.346 | 0.358 | 0.404 |
| GBLUP: SVM-pol | 2.201 | 1.464 | 1.040 | 1.254 | 1.486 | 1.380 | 1.335 |
| GBLUP: SVM-sig | 1.118 | 1.126 | 0.765 | 0.753 | 0.915 | 0.832 | 0.788 |
| SVM-lin: SVM-rad | 2.337 | 1.370 | 0.873 | 0.910 | 0.886 | 0.733 | 0.564 |
| SVM-lin: SVM-pol | 0.400 | 0.183 | 0.010 | 0.200 | 0.513 | 0.391 | 0.497 |
| SVM-lin: SVM-sig | 1.431 | 0.505 | 0.300 | 0.380 | 0.286 | 0.252 | 0.161 |
| SVM-rad: SVM-pol | 1.939 | 1.168 | 0.862 | 1.043 | 1.218 | 1.051 | 1.000 |
| SVM-rad: SVM-sig | 0.882 | 0.826 | 0.566 | 0.528 | 0.584 | 0.475 | 0.400 |
| SVM-pol: SVM-sig | 1.036 | 0.321 | 0.303 | 0.551 | 0.733 | 0.614 | 0.638 |

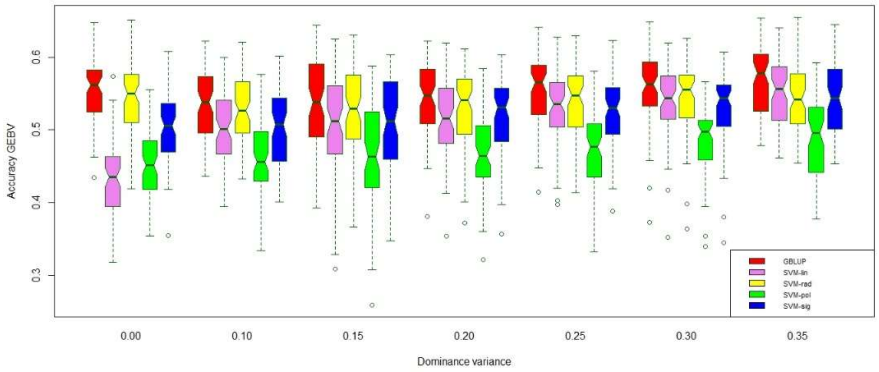


Fig. 1. Prediction accuracy of GEBV of GBLUP and support vector machine using different linear (SVM-lin), radial (SVM-rad), polynomial (SVM-pol), and cyclic (SVM-sig) kernel functions at different levels of dominance variance

شکل ۱- صحت پیش‌بینی GEBV روش‌های GBLUP و ماشین بردار پشتیبان بااستفاده از توابع کرنل مختلف خطی (SVM-lin)، شعاعی (SVM-rad)، چندجمله ای (SVM-pol) و حلقوی (SVM-sig) در سطوح مختلف واریانس غالبیت

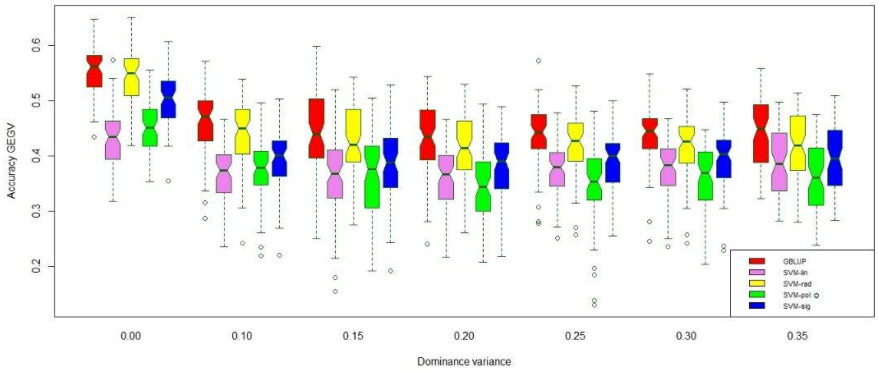


Fig. 2. Prediction accuracy of GEGV of GBLUP and support vector machine using different linear (SVM-lin), radial (SVM-rad), polynomial (SVM-pol), and cyclic (SVM-sig) kernel functions at different levels of dominance variance

شکل ۲- صحت پیش‌بینی GEGV روش‌های GBLUP و ماشین بردار پشتیبان با استفاده از توابع کرنل مختلف خطی (SVM-lin)، شعاعی (SVM-rad)، چندجمله ای (SVM-pol) و حلقوی (SVM-sig) در سطوح مختلف واریانس غالبیت

گزارش شده است که SVM مبتنی بر شعاعی در مقایسه با SVM مبتنی بر سیگموئید، عملکرد پیش‌بینی نسبتاً بالاتری دارد (Kasnavi *et al.,* 2018)، اگرچه تفاوت‌ها قابل توجه نبود، که این نتیجه با پژوهش حاضر اندکی مغایرت دارد، زیرا زمانی‌که واریانس غالبیت صفر بود (مدل صرفاً افزایشی)، عملکرد SVM-rad در مقابل SVM-sig برتری قابل توجهی داشت. که این مغایرت را می توان به نوع (پیوسته یا گسسته) و معماری مختلف ژنتیکی صفات مورد تجزیه نسبت داد، هر چند که در مدل افزایشی-انحراف غالبیت با افزایش واریانس غالبیت، این برتری کاهش یافت. در یک مدل ژنتیکی صرفاً افزایشی، رابطه بین نشانگرها و صفت غالباً خطی یا به‌صورت جزئی غیرخطی است. تابع کرنل شعاعی برای چنین سناریوهایی عملکرد بسیار مطلوبی دارد، زیرا می‌تواند روابط خطی را به‌طور موثر برآورد نماید و در عین حال در برابر نویز بسیار مقاوم است (Gianola *et al.,* 2006; Gonzalez-Recio *et al.,* 2014). تابع کرنل سیگموئید از شبکه‌های عصبی الهام گرفته شده است و معمولاً برای حل مسائل طبقه‌بندی دو دسته‌ای استفاده می‌شود. این توابع در مسائل رگرسیون مانند پیش‌بینی ژنومی کمتر مورد استفاده قرار می‌گیرند، زیرا یک رابطه $S$ شکل خاص را فرض می‌کند که ممکن است با معماری ژنتیکی افزایشی همسو نباشد (Hastie *et al.,* 2009). در یک مدل ژنتیکی صرفاً افزایشی، آثار مکان‌های کروموزومی به‌صورت خطی و مستقل فرض می‌شوند. تابع کرنل شعاعی به‌دلیل توانایی آن در تقریب روابط خطی در حالی‌که انعطاف‌پذیری خود را حفظ می‌کند، با این فرض هماهنگ است (Gianola *et al.,* 2006)، در حالی که تابع کرنل سیگموئید از یک ساختار غیرخطی بهره می‌برد که ممکن است با معماری ژنتیکی افزایشی مطابقت نداشته باشد و منجر به عملکرد کمتر از حد مطلوب شود (Hastie *et al.,* 2009). توابع کرنل شعاعی دارای فراپارامترهای کمتری برای تنظیم در مقایسه با توابع سیگموئید هستند که بهینه‌سازی آن را برای کارهای پیش‌بینی ژنومی آسان‌تر می‌کند. این سادگی اغلب منجر به تعمیم‌پذیری بهتری نیز می‌شود (Gonzalez-Recio *et al.,* 2014). توابع کرنل سیگموئید به‌شدت به انتخاب فراپارامترها بستگی دارند و تنظیم نادرست، به‌ویژه در مجموعه داده‌هایی با معماری ژنتیکی افزایشی، می‌تواند منجر به برازش بیش از حد یا عدم برازش شود (Scholkopf and Smola, 2002).

مطالعات تجربی در پیش‌بینی ژنومی نیز نشان داده‌اند که کرنل شعاعی از سایر کرنل‌ها، از جمله کرنل سیگموئید، برای صفات دارای ساختارهای ژنتیکی افزایشی بهتر عمل می‌کند (Gianola *et al.,* 2006; Gonzalez-Recio *et al.,* 2014). کرنل سیگموئید، در حالی که برای برخی از شرایط مفید است (برای مثال، طبقه‌بندی دو دسته‌ای یا روابط شدید غیرخطی)، به‌طور کلی برای مدل‌های ژنتیکی افزایشی کمتر موثر است (Hastie *et al.,* 2009). در پژوهش حاضر، عملکرد روش پارامتری GBLUP نسبت به روش SVM-rad بالاتر بود. GBLUP به‌طور خاص برای آثار ژنتیکی افزایشی، که اغلب منبع اصلی تنوع ژنتیکی در بسیاری از صفات هستند، طراحی شده است. این روش فرض می‌نماید که آثار ژنتیکی دارای توزیع نرمال بوده و می‌توان با استفاده از ماتریس رابطه ژنومی مدل‌سازی کرد (VanRaden, 2008). اگر صفت مورد مطالعه عمدتاً با آثار ژنتیکی افزایشی کنترل شود، انتظار می‌رود GBLUP عملکرد خوبی داشته باشد زیرا مستقیماً این آثار را با استفاده از یک مدل مختلط خطی مدل‌سازی می‌کند (Meuwissen *et al.,* 2001). در مقابل، توابع کرنل شعاعی، در حالی‌که انعطاف‌پذیر هستند، ممکن است در صورتی که ساختار ژنتیکی عمدتاً افزایشی باشد، داده‌ها را بیش از حد یا کمتر از حد برازش کنند، زیرا به‌صراحت روابط افزایشی را مدل‌سازی نمی‌کنند. GBLUP یک روش پارامتری است که بر مفروضات آماری تثبیت شده (مثلاً نرمال بودن آثار ژنتیکی) متکی است و برآوردهای قابل تفسیری از فراسنجه‌های ژنتیکی مانند وراثت‌پذیری و ارزش‌های اصلاحی ارائه می‌دهد (VanRaden, 2008). تابع کرنل شعاعی، از سوی دیگر، یک روش ناپارامتری یادگیری ماشین است که بر فرضیات صریح در مورد توزیع آثار ژنتیکی متکی نیست. در حالی که این انعطاف‌پذیری می‌تواند در روابط پیچیده غیرخطی مفید باشد، اما ممکن است زمانی که معماری ژنتیکی عمدتاً افزایشی و خطی باشد، منجر به کاهش عملکرد شود (Gianola *et al.,* 2006). GBLUP به‌ویژه در سناریوهایی که تعداد نشانگرها نسبت به تعداد افراد زیاد است مؤثر است، زیرا از یک ماتریس رابطه ژنومی برای جمع‌بندی اطلاعات نشانگرها استفاده می‌کند (VanRaden, 2008). در این مورد، تابع کرنل شعاعی ممکن است با داده‌های ژنومی با ابعاد بالا، دچار مشکل شود، زیرا به تنظیم دقیق فراپارامترها (برای

مثال، فراسنجه تنظیم و فراسنجه کرنل برای جلوگیری از برازش یا بیش از حد یا کمتر از حد) نیاز دارد (Scholkopf and Smola, 2002). چندین مطالعه نشان داده‌اند که GBLUP اغلب از روش‌های یادگیری ماشین مانند SVM برای صفات با ساختارهای ژنتیکی عمدتاً افزایشی بهتر عمل می‌کند (Gianola *et al.,* 2006; Gonzalez-Recio *et al.,* 2014). در پژوهش پیشین، نشان داده شده است که روش‌های یادگیری ماشین مانند تابع کرنل شعاعی زمانی‌که معماری ژنتیکی شامل آثار غیرافزایشی قابل توجهی باشد (به‌عنوان مثال، غالبیت یا اپیستازی) یا زمانی که رابطه بین نشانگرها و صفت بسیار غیرخطی است، بهتر عمل می‌کنند (Howard *et al.,* 2014)، که با پژوهش حاضر مغایرت دارد. علت این مغایرت را می‌توان به تفاوت در مولفه‌های ژنتیکی تشکیل‌دهنده فنوتیپ صفت مورد مطالعه نسبت داد. همچنین، تابع کرنل مانند شعاعی و سایر مدل‌های پیچیده معمولاً به اندازه‌های نمونه بزرگ‌تری برای ثبت الگوهای پیچیده نیاز دارند. در شبیه‌سازی حاضر، با شش کروموزوم، مجموعه داده نسبتاً کوچک است، و ممکن است که روش کرنل بیش از حد برازش یابد، که منجر به عملکرد ضعیف‌تر در مقایسه با مدل‌های ساده‌تر می‌شود. در پژوهشی دیگر، بر اساس مدل صرفاً افزایشی گزارش شد که روش‌های پارامتری نسبت به روش‌های ناپارامتری SVM براساس تابع شعاعی و جنگل تصادفی، عملکرد بهتری دارند. همچنین، نشان داده شد که روش SVM می‌تواند صحت پیش‌بینی بالاتری نسبت به روش جنگل تصادفی ارائه دهد (Sahebalam *et al.,* 2019). در پژوهشی دیگر، از روش‌های غیرپارامتری برای رتبه‌بندی افراد بر اساس ارزش اصلاحی ژنومی استفاده شد و گزارش شد زمانی که SVM بر اساس هسته شعاعی ساخته شد، حداکثر صحت به‌دست آمد (Blondel *et al.,* 2015). مطالعه دیگری بر اساس تجزیه و تحلیل داده‌های یک جمعیت ناهمگن موش، عملکرد پیش‌بینی ۱۰ روش آماری مختلف به‌کار گرفته شده در انتخاب ژنومی را مورد مقایسه قرار داد و پیشنهاد شد که یک SVM مبتنی بر شعاعی نسبت به سایر روش‌ها، به‌ویژه در ارزیابی ژنومی صفات با وراثت‌پذیری بالا، برتری دارد (Neves *et al.,* 2012). پژوهشی نشان داد که ساخت کرنل شعاعی به‌دلیل تعداد فراسنجه‌های کمتر، آسان‌تر از کرنل چندجمله‌ای است (یک در مقابل سه فراسنجه). همچنین، خاطر نشان شد که ارزش‌های بی‌اعتباری در برخی از فراسنجه‌های کرنل سیگموئید مورد انتظار است (Zhu *et al.,* 2010).

در پژوهش حاضر، استفاده از مدل افزایشی-انحراف غالبیت به‌خصوص زمانی‌که واریانس غالبیت سهم بیشتری در واریانس فنوتیپی داشت صحت پیش‌بینی GEBV افزایش یافت، زیرا در سطح ژن، انحراف غالبیت از برهمکنش بین آلل‌ها در یک مکان ژنی ناشی می‌شود. آثار افزایشی ژنوتیپ‌ها در یک مکان به‌صورت ارزش‌های اصلاحی بیان می‌شوند که بخشی از آثار غالبیت را شامل می‌شود، زیرا افراد آلل‌ها را به فرزندان خود منتقل می‌کنند نه ژنوتیپ‌ها را. در نظر گرفتن آثار غالبیت در ارزیابی ژنومی می‌تواند منجربه افزایش صحت پیش‌بینی GEBV شود. پژوهشی که در رابطه با تجزیه هفت صفت اقتصادی گوسفند مغانی بر-اساس دو مدل افزایشی و افزایشی-غالبیت انجام شد پیشنهاد کرد زمانی‌که بخش بزرگی از تنوع فنوتیپی صفات با آثار غالبیت توجیه شود، مدل افزایشی-غالبیت نسبت به مدل افزایشی مزیت دارد و باعث بهبود صحت پیش‌بینی ژنومی می‌شود (Seyedsharifi *et al.,* 2022). در تحقیقی دیگر، ارزیابی ژنومی در مورد صفات شیر تولیدی، مقدار چربی و پروتئین در گاوهای شیری هلشتاین ایران صورت گرفت. با وجود سهم اندک آثار غالبیت در تنوع فنوتیپی این صفات، لحاظ نمودن اثر غالبیت در مدل منجر به افزایش صحت پیش‌بینی ژنومی نسبت به مدل صرفاً افزایشی شد (Mohammadi, 2019). در پژوهشی دیگر با استفاده از شبیه‌سازی رایانه‌ای گزارش شد در صورتی که آثار ژنتیکی غالبیت در تنوع فنوتیپی صفت مشارکت داشته باشند، اما در مدل ارزیابی ژنومی لحاظ نشده و به‌صورت تفکیک نشده از آثار افزایشی باقی بمانند، منجر به کاهش صحت ارزش‌های اصلاحی تا حدود ۲۵٪ می‌شود. همچنین، قابلیت اعتماد ارزش‌های اصلاحی ژنومی با افزایش درصد QTLهای دارای اثر غالبیت تا حدود ۴۰٪ کاهش یافت (Karimi *et al.,* 2023). در پژوهشی دیگر، با استفاده از صفت شبیه‌سازی شده با ۲۰ QTL، نشان داده شد که کنش ژنی غالبیت و اپیستازی، صحت پیش‌بینی روش‌های ارزیابی ژنومی را کاهش می‌دهد (Salehi *et al.,* 2020). مطالعات نشان داد زمانی‌که عملکرد ژن پیچیده‌تر بود صحت پیش-بینی هر دو روش پارامتری و ناپارامتری کاهش یافت (Momen *et al.,* 2018). نتایج این دو پژوهش با پژوهش حاضر بر اساس صحت پیش‌بینی GEGV همخوانی دارد.

## نتیجه‌گیری کلی

با عنایت به نتایج حاصل از این پژوهش، می‌توان اذعان داشت که رویکرد GBLUP، به سبب ماهیت پارامتری خود و توانایی ذاتی در مدل‌سازی آثار ژنتیکی افزایشی، صحت پیش‌بینی بالاتری را برای GEBV و GEGV ارائه نمود. همچنین، می‌توان بیان کرد، در عدم حضور واریانس غالبیت، روش SVM-rad عملکرد مطلوبی دارد. با این وجود، شایان ذکر است که با افزایش واریانس غالبیت، کارآیی روش‌های مبتنی بر ماشین بردار پشتیبان، به‌ویژه کرنل خطی (SVM-lin) بهبود چشمگیری یافته و شکاف عملکردی با GBLUP کاهش می‌یابد. این مهم، حاکی از ظرفیت بالقوه SVM در بررسی آثار غیرافزایشی، به‌ویژه در شرایطی است که سهم غالبیت در تبیین واریانس فنوتیپی افزایش می‌یابد. در مجموع، نتایج این مطالعه، ضمن تأیید برتری روش GBLUP در شرایط حاکمیت آثار افزایشی، بر اهمیت انتخاب کرنل مناسب در مدل ناپارامتری SVM، متناسب با معماری ژنتیکی صفت مورد مطالعه، تأکید می‌نماید.

## فهرست منابع


Akbarpour, T., Ghavi Hossein-Zadeh, N., & Shadparvar, A. A. (2021). Marker genotyping error effects on genomic predictions under different genetic architectures. *Molecular Genetics and Genomics*, *296*, 79-89. doi: 10.1007/s00438-020-01728-z.

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., & Hayes, B. J. (2016). Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. *Genetics Selection Evolution*, *48*, 1-11. doi: 10.1186/s12711-016-0186-0

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., Goddard, M. E., & Hayes, B. J. (2017). Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. *Journal of Dairy Science*, *100*(2), 1203-1222. doi: 10.3168/jds.2016-11261

Ansari, S., Ghavi Hossein-Zadeh, N., & Shadparvar, A. A. (2024). Genomic predictions under different genetic architectures are impacted by mating designs. *Veterinary and Animal Science*, *25*, 100373. doi: 10.1016/j.vas.2024.100373

Atefi, A., Shadparvar, A. A., & Hossein-Zadeh, N. G. (2021). Accuracy of genomic evaluation considering the interaction effect between estimation method of marker effects, population structure, and genetic architecture of the trait. *Animal Production Research*, *10*(2), 1-10. doi: 10.22124/ar.2021.16234.1520 [In Persian]

Blondel, M., Onogi, A., Iwata, H., & Ueda, N. (2015). A ranking approach to genomic selection. *PloS One*, *10*(6), e0128570. doi: 10.1371/journal.pone.0128570

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational Learning Theory*. Pp. 144-152. doi: 10.1145/130385.130401

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. New York, NY: Routledge Academic. doi: 10.4324/9780203771587

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*, 273-297. doi: 10.1007/BF00994018.

Crow, J. F. (2010). On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1544), 1241-1244. doi: 10.1098/rstb.2009.0275

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*(1), 375-385. doi: 10.1534/genetics.109.101501

Duenk, P., Calus, M. P., Wientjes, Y. C., & Bijma, P. (2017). Benefits of dominance over additive models for the estimation of average effects in the presence of dominance. *G3: Genes, Genomes, Genetics*, *7*(10), 3405-3414. doi: 10.1534/g3.117.300113

Esfandyari, H., & Sørensen, A. C. (2017). xbreed: an R package for genomic simulation of purebreds and crossbreds. In *Book of Abstracts of the 68th Annual Meeting of the European Federation of Animal Science*. Pp. 234-234. doi: 10.3920/9789086868599_313

Falconer, D. S., and McKay, T. (1996). Introduction to quantitative genetics. Harlow: Pearson Education Limited.

Gianola, D., Fernando, R. L., & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, *173*(3), 1761-1776. doi: 10.1534/genetics.105.049510

Ghafouri-Kesbi, F., Rahimi-Mianji, G., Honarvar, M., & Nejati-Javaremi, A. (2016). Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. *Animal Production Science*, *57*(2), 229-236. doi: 10.1071/AN15538

Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, *124*(6), 323-330. doi: 10.1111/j.1439-0388.2007. 00702.x

González-Recio, O., Rosa, G. J., & Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science*, *166*, 217-231. doi: 10.1016/j.livsci.2014.05.036

Habier, D., Fernando, R. L., & Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*(4), 2389-2397. doi: 10.1534/genetics.107.081190

Hastie, T. J., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning, 2nd edition. Springer-Verlag, New York.

Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, *91*(1), 47-60. doi: 10.1017/S0016672308009981

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, *32*(1), 69-83. doi: 10.2307/2529339

Hill, W. G. (2010). Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1537), 73-85. doi: 10.1098/rstb.2009.0203

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, *38*, 226-231. doi: 10.1007/BF01245622

Howard, R., Carriquiry, A. L., & Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes, Genomes, Genetics*, *4*(6), 1027-1046. doi: 10.1534/g3.114.010298

Karimi, M., Ghafouri-Kesbi, F., & Zamani, P. (2023). Investigating the impact of dominance genetic effects on the accuracy of genomic evaluation. *Research on Animal Production*. *14*(1), 145-153. doi:10.61186/rap.14.39.145 [In Persian]

Kasnavi, S. A., Aminafshar, M., Shariati, M. M., Kashan, N. E. J., & Honarvar, M. (2018). The effect of kernel selection on genome wide prediction of discrete traits by Support Vector Machine. *Gene Reports*, *11*, 279-282. doi: 10.1016/j.genrep.2018.04.006

Long, N., Gianola, D., Rosa, G. J., & Weigel, K. A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. *Theoretical and Applied Genetics*, *123*, 1065-1074. doi: 10.1007/s00122-011-1648-y

Mäki-Tanila, A. (2007). An overview on quantitative and genomic tools for utilising dominance genetic variation in improving animal production. *Agricultural and Food Science*, *16*(2), 188-198. doi: 10.2137/145960607782219337

Meuwissen, T. H., Hayes, B. J., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819-1829. doi: 10.1093/genetics/157.4.1819

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., & Lin, C. C. (2019). e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. Available at: https://cran.r-project.org/web/packages/e1071/e1071.pdf

Mohammadi, Y. (2019). Accuracy of genomic selection using models with additive effects for productive traits in Iranian Holstein cows. The second international conference and the third national conference on agriculture, environment and food security. Jiroft University, Jiroft, Iran. [In Persian]

Momen, M., Mehrgardi, A. A., Sheikhi, A., Kranis, A., Tusell, L., Morota, G., Rosa, G. J. M., & Gianola, D. (2018). Predictive ability of genome-assisted statistical models under various forms of gene action. *Scientific Reports*, *8*(1), 12309. doi: 10.1038/s41598-018-30089-2

Neves, H. H., Carvalheiro, R., & Queiroz, S. A. (2012). A comparison of statistical methods for genomic selection in a mice population. *BMC Genetics*, *13*, 1-17. doi: 10.1186/1471-2156-13-100

Nocedal, J., & Wright, S. J. (Eds.). (1999). Numerical optimization. New York, NY: Springer New York. doi: 10.1007/0-387-22742-3_18

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*, 1-5. doi: 10.1186/1753-6561-5-S3-S11

Pérez, P., & de Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483-495. doi: 10.1534/genetics.114.164442

Quaas, R. L. (1976). Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics*, *32*(4), 949-953. doi: 10.2307/2529279

Saheb Alam, H., Gholizadeh, M., Hafezian, H., & Farhadi, A. (2018). Comparison of Bayesian methods in the genomic evaluation with different genetic architecture. *Research on Animal Production*, *8*(18), 177-186. doi: 10.29252/rap.8.18.177 [In Persian]

Sahebalam, H., Gholizadeh, M., Hafezian, H., & Farhadi, A. (2019). Comparison of parametric, semiparametric and nonparametric methods in genomic evaluation. *Journal of Genetics*, *98*, 1-8. doi: 10.1007/s12041-019-1149-3

Sahebalam, H., Gholizadeh, M., Hafezian, H., & Ebrahimi, F. (2022). Evaluation of Bagging approach versus GBLUP and Bayesian LASSO in genomic prediction. *Journal of Genetics*, *101*(1), 19. doi: 10.1007/s12041-022-01358-x

Sahebalam, H., Gholizadeh, M., & Hafezian, H. (2024). Investigating the performance of frequentist and Bayesian techniques in genomic evaluation. *Biochemical Genetics*, 1-27. doi: 10.1007/s10528-024-10842-1

Salehi, A., Bazrafshan, M., & Abdollahi-Arpanahi, R. (2021). Assessment of parametric and non-parametric methods for prediction of quantitative traits with non-additive genetic architecture. *Annals of Animal Science*, *21*(2), 469-484. doi: 10.2478/aoas-2020-0087

Schölkopf, B. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond.

Seyedsharifi, R., Ala Noshahr, F., Seif Davati, J., & Hedayat Evrigh, N. (2022). Genomic prediction of additive and dominance effects on some economic traits of Moghani sheep. *Research on Animal Production*. *13*(38), 187-193. doi:10.52547/rap.13.38.187 [In Persian]

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, *28*(1), 127-135. doi: 10.1016/j.eswa.2004.08.009

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., & Lund, M. S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One*. *7*, e45293. doi: 10.1371/journal.pone.0045293

Tamaddoni-Arani, M., Razmkabir, M., Abdollahi-Arpanahi, R., Rashidi, A., & Moradi, Z. (2021). Comparison of different statistical methods in genomic selection based on selection effectiveness criteria. *Animal Production Research*, *10*(3), 1-20. doi: 10.22124/ar.2021.19332.1608 [In Persian]

Thomasen, J. R., Sørensen, A. C., Su, G., Madsen, P., Lund, M. S., & Guldbrandtsen, B. (2013). The admixed population structure in Danish Jersey dairy cattle challenges accurate genomic predictions. *Journal of Animal Science*, *91*(7), 3105-3112. doi: 10.2527/jas.2012-5490

Toro, M. A., & Varona, L. (2010). A note on mate allocation for dominance handling in genomic selection. *Genetics Selection Evolution*, *42*, 1-9. doi: 10.1186/1297-9686-42-33

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414-4423. doi: 10.3168/jds.2007-0980

Vapnik, V. (1995). The nature of statistical learning theory. (2nd ed.). Springer. doi: 10.1007/978-1-4757-3264-1

Varona, L., Legarra, A., Toro, M. A., & Vitezica, Z. G. (2018). Non-additive effects in genomic selection. *Frontiers in Genetics*, *9*(78), 1-12. doi: 10.3389/fgene.2018.00078

Vitezica, Z. G., Varona, L., & Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, *195*(4), 1223-1230. doi: 10.1534/genetics.113.155176

Yang, P., Hwa Yang, Y., Zhou, B. B., & Zomaya, A. Y. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, *5*(4), 296-308. doi: 10.2174/157489310794072508

Zeng, J., Toosi, A., Fernando, R. L., Dekkers, J. C., & Garrick, D. J. (2013). Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genetics Selection Evolution*, *45*, 1-17. doi: 10.1186/1297-9686-45-11

Zhu, Y., Tan, Y., Hua, Y., Wang, M., Zhang, G., & Zhang, J. (2010). Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. *Journal of Digital Imaging*, *23*, 51-65. doi: 10.1007/s10278-009-9185-9