Identification of genes affecting growth traits in broiler chickens using linear regression and machine learning methods

ABSTRACT

Knowledge about the association between single nucleotide polymorphisms (SNPs) and important economic traits is one of the crucial tools in breeding programs within the poultry industry. Genome-wide studies for discovering SNP variations related to these traits are often conducted using simple linear models. However, due to certain assumptions of these models, some SNP markers may not be identified. This study aimed to evaluate the performance of random forest and gradient boosting methods compared to linear models in identifying SNP markers associated with body weight traits at 6 and 9 weeks of age in F2 broiler chickens resulting from crosses between the commercial Arian line and native Urmia birds. The results showed that the machine learning approaches were able to identify important markers, such as GGaluGA308573, GGaluGA255033, Gga_rs13614212, Gga_rs13743072, GGaluGA258772, Gga_rs14034395, and Gga_rs13858398, associated with body weight traits, which were related to genes *MAP2*, *ACSL1*, *CAMSAP2*, *FAM117B*, *SLC4A4*, *TIMP4*, *and LncRNA*, respectively. These genes are primarily involved in cellular division, growth control, regulation of cellular skeleton structure and microtubules, and transcription activity, constituting the most important biological processes. The identification of these novel genes using machine learning methods, which were not detected by linear models and previous studies in this population, could provide new insights into genetic control of growth traits in broiler chickens. Moreover, the discovered significant markers can be utilized in genetic improvement programs for birds.

Keywords: *nucleotide polymorphisms, Genome-wide studies, broiler chickens, machine learning.*

شناسایی ژن‌های موثر بر صفات رشد در جوجه های گوشتی با استفاده از روش‌های رگرسیون خطی و یادگیری ماشین

چکیده

آگاهی از ارتباط چندشکلی‌های تک‌نوکلئوتیدی با صفات مهم اقتصادی یکی از ابزارهای مهم برنامه‌های اصلاح نژاد در صنعت طیور است. مطالعات پویش ژنومی برای کشف چندشکلی‌های تک‌نوکلئوتیدی (نشان‌گرها) مرتبط با این صفات، اغلب با استفاده از مدل‌های خطی ساده صورت می‌گیرد که به دلیل وجود برخی از فرضیات این مدل‌ها، ممکن است بعضی از نشان‌گرها شناسایی نشوند. این مطالعه با هدف ارزیابی کارآیی روش‌های جنگل تصادفی و گرادیان بوستینگ و ارزیابی عملکرد آن‌ها در مقابل مدل خطی برای شناسایی نشان‌گرهای همبسته با صفات وزن بدن در سنین ۶ و ۹ هفتگی در جوجه‌های گوشتی نسل دوم حاصل از تلاقی‌های دوطرفه لاین تجاری آرین با پرنده‌های بومی ارومیه انجام شد. نتایج نشان داد که دو روش یادگیری ماشین توانستند نشان‌گرهای مهمی از جمله GGaluGA308573، GGaluGA255033، Gga_rs13614212، Gga_rs13743072، GGaluGA258772، Gga_rs14034395 و Gga_rs13858398 را برای صفات وزن بدن شناسایی کنند که به ترتیب با ژن‌های *MAP2*، *ACSL1*، *CAMSAP2*، *FAM117B*، *SLC4A4*، *TIMP4* و *LncRNA* در ارتباط بودند. تقسیم سلولی، کنترل رشد، تنظیم ساختار اسکلت سلولی و میکروتوبول، و فعالیت رونویسی مهمترین فرآیند بیولوژیکی این ژن‌ها می‌باشند. مطالعه ژن‌های جدید شناسایی شده توسط روش‌های یادگیری ماشین، که مدل خطی قادر به شناسایی آن‌ها در جمعیت مورد مطالعه نبودند، می‌تواند بینش جدیدی را برای کنترل ژنتیکی صفات رشد در جوجه‌های گوشتی باز کند. علاوه بر این، نشان‌گرهای با اهمیت کشف شده، قابلیت استفاده در برنامه‌های اصلاح ژنتیکی پرندگان را دارند.

کلیدواژه‌ها: *چندشکلی‌های تک‌نوکلئوتیدی ، مطالعات پویش ژنومی، جوجه های گوشتی، یادگیری ماشین.*

مقدمه و پیشینه پژوهش

در پرورش جوجه‌های گوشتی، صفت وزن بدن یک صفت مهم اقتصادی و پیچیده و از صفات بسیار مهم و تاثیرگذار در علم پرورش طیور است (Mebratie *et al*., 2019). برنامه‌های اصلاح‌نژاد برای افزایش سرعت رشد از طریق انتخاب در لاین‌های تجاری باعث کوتاه‌تر شدن دوره پرورش جوجه‌های گوشتی شده است. در گذشته به دلیل در دسترس نبودن چندشکلی‌های تک‌نوکلئوتیدی (Single nucleotide polymorphism, SNP) موثر بر تفاوت‌های ژنتیکی بین افراد یا بین جمعیت‌ها، اصلاح طیور بدون داشتن دانشی از ژن‌های اساسی موثر بر صفات تحت انتخاب صورت می‌گرفت. اما با به‌کارگیری دانش مطالعات پویش ژنومی (Genome-Wide Association Studies, GWAS) و شناسایی نشان‌گرها و ارتباط آن‌ها با ژن‌های موثر بر صفات، اطلاعات مولکولی لازم برای انتخاب به کمک نشان‌گرها و ژن‌ها برای بهبود صفات کمی فراهم شده است (Goddard and Hayes, 2009).

در مطالعات پویش ژنوم از وجود تعداد زیادی چندشکلی در سطح ژنوم به همراه فنوتیپ و اطلاعات شجره جهت تجزیه همبستگی بین ژنوتیپ و فنوتیپ استفاده و ژن‌ها و عناصر تنظیم کننده مرتبط با صفات مهم اقتصادی شناسایی می شود (Hayes, 2013). در این مطالعات، رایج‌ترین مبنا برای بررسی اثر نشان‌گرها، استفاده از روش تابعیت خطی (Linear Model) و P-value است. با وجود گزارشات متعدد در باره‌ی مطالعات پویش ژنومی صفات رشد در جوجه‌های آمیخته گوشتی (Zhang *et al*., 2015; Emrani *et al*., 2017; Cha *et al*., 2021; Dadousis *et al*., 2021)، در بررسی‌های مختلف به مشکلات استفاده از این روش، مانند بالا بردن نرخ خطای نوع اول و بیش برآورد شدن اثر نشان‌گرها، در نظر نگرفتن عدم تعادل پیوستگی بین نشان‌گرها، فرض نادرست مستقل بودن نشان‌گرها از یک‌دیگر، پیروی‌کردن تمام متغیرهای ژنومی از یک توزیع نرمال و در نظر نگرفتن اثر متقابل بین نشان‌گرها اشاره شده است (Wray *et al*., 2013; Sun *et al*., 2021, Enoma *et al*., 2022). در نتیجه، یک جایگزین مناسب برای حل این مشکلات، انجام مطالعات پویش ژنومی بر مبنای روش‌های یادگیری ماشین است.

روش‌های یادگیری ماشین به عنوان یک ابزار برای بهینه‌سازی توان پیش‌بینی در مجموعه‌ی داده‌های حجیم برای مطالعات علوم دامی به طور گسترده‌ای در مطالعات پویش ژنومی استفاده شده است (Li *et al*., 2018; Hu *et al* 2020). این روش‌ها،

الگوریتم‌های یادگیری داده محور هستند که می‌توانند با استفاده از داده‌های موجود، آموزش‌دیده و از الگوی کشف شده، در پیش‌بینی سناریوهای آینده استفاده کنند. یادگیری ماشین بسته به نوع آموزش، می‌تواند تحت نظارت یا بدون نظارت باشد. یادگیری نظارت شده از یک مجموعه‌ی آموزشی برای آموزش مدل‌ها و به دست آوردن خروجی مورد نظر استفاده می‌کند. این مجموعه‌ی آموزشی، شامل ورودی‌ها و خروجی‌های صحیح است که به مدل اجازه می دهد در طول زمان آموزش ببیند (Baştanlar and Özuysal, 2014). در میان روش‌های یادگیری نظارت شده می توان به روش جنگل تصادفی (Random Forest, RF) و گرادیان بوستینگ (Gradient Boosting, GB) اشاره نمود.

مطالعات نشان داده است که روش‌های جنگل تصادفی و گرادیان بوستینگ در مطالعات پویش ژنومی، روش‌های مناسبی برای انتخاب نشان‌گرهای مهم و شناسایی ژن‌های تاثیرگذار بر صفات اقتصادی هستند. این روش‌ها بر خلاف مدل‌های آماری دیگر، نیازی به تعیین توزیع جمعیت مورد بررسی، ارزیابی برآوردگرها (فاصله اطمینان یا p-value) یا آزمون فرضیه‌های صفر ندارند و روش آزاد و بدون آمار هستند. علاوه بر این، این روش‌ها قادر هستند نشان‌گرهایی بدون تاثیر و یا دارای تاثیر منفی بر صفت را مشخص کنند و روابط غیرخطی بین نشان‌گرها و فنوتیپ را در نظر بگیرند و در داده‌هایی با ابعاد بزرگ به راحتی قابل استفاده باشند (Li *et al*., 2018; Sun *et al*., 2022).

عواملی مانند عدم تعادل پیوستگی، اثرات غیرافزایشی، فراوانی آللی جزئی، روابط پیچیده و برهم‌کنش بین نشان‌گرها بر عملکرد روش‌های RF و GB موثر باشند (Li *et al*., 2018; Enoma *et al*., 2022). با این حال، مطالعات انجام شده برای بررسی الگوریتم‌های مختلف یادگیری ماشین برای مطالعات پویش ژنومی در جوجه‌های گوشتی بسیار محدود است. بنابراین، هدف از مطالعه‌ی حاضر، شناسایی نشان‌گرهای مهم و تاثیرگذار بر صفات وزن بدن جوجه‌های نسل F_2 با استفاده از دو روش RF و GB و مقایسه آن‌ها با روش مدل خطی (Linear Model, LM) است.

روش شناسی پژوهش

داده‌های فنوتیپی استفاده شده در پژوهش حاضر، از پرنده‌های آمیخته نسل دوم (F_2) حاصل از تلاقی دو طرفه‌ی بین پرنده‌های لاین تجاری آرین (A) و بومی استان آذربایجان غربی (U)، که در مرکز تحقیقات پرورش طیور دانشکده کشاورزی دانشگاه تربیت مدرس ایجاد شدند، بود. پرندگان F_1 از آمیزش ♀ U × ♂ A و ♀ A × ♂ U ایجاد شدند. به منظور ایجاد پرنده‌های نسل دوم، نرهای F_1 از هر تلاقی متقابل یا معکوس، هر کدام با ۴ تا ۶ قطعه از ماده‌های خانواده‌های دیگر آمیزش داده شدند. تخم‌مرغ‌های جمع‌آوری شده، ابتدا شماره‌گذاری شده، سپس در دمای ۱۲ درجه سانتی‌گراد نگهداری شدند. پس از جداسازی تخم‌مرغ های کثیف، بدشکل و نامناسب، بقیه تخم مرغ‌ها جهت جوجه‌کشی در سینی‌های ستر قرار و سپس در روز ۱۸ به هچر انتقال داده شدند. جوجه‌ها پس از تولد تا ۷ روزگی به صورت گروهی نگهداری و به وسیله دان‌خوری سینی و آبخوری کله‌قندی تغذیه شدند و از ۸ روزگی تا پایان دوره به صورت انفرادی در قفس‌های تکی پرورش یافتند. این پرنده‌ها با استفاده از سه جیره غذایی پلت شده شامل، آغازین، رشد و پایانی پرورش داده شدند. رکوردهای وزن بدن از تولد تا پایان دوره پرورش به‌صورت هفتگی ثبت گردید. برای مطالعه حاضر، از وزن بدن ثبت شده در سنین ۶ هفته (BW6) و ۹ هفته (BW9) استفاده شد. آماره‌های توصیفی صفات مورد مطالعه در جدول ۱ نشان داده شده است.

جدول ۱. آماره توصیفی صفات وزن بدن در ۶ هفتگی (BW6) و ۹ هفتگی (BW9)

صفت	تعداد	میانگین	انحراف معیار	حداقل	حداکثر
BW6	۳۰۰	۱۳۲۶	۲۵۷	۷۷۰	۲۰۵۰
BW9	۲۹۸	۲۲۹۴	۳۲۷	۱۷۵۵	۳۰۸۰

برای پرنده‌های تحقیق حاضر، نمونه‌های خون از جوجه‌های نسل دوم جمع‌آوری شد. استخراج DNA با استفاده از روش بهینه رسوب‌دهی نمکی انجام گرفت. جهت تعیین ژنوتیپ هر یک از افراد، نمونه‌ها به دانشگاه Aarhus کشور دانمارک انتقال داده شدند. در مجموع، تعداد ۳۱۲ نمونه DNA از نسل F_2 با همکاری این دانشگاه با استفاده از تراشه SNP تجاری ژنوم مرغ بنام Illumina 60K SNP chip (Illumnia industry, USA)، که توسط شرکت تجاری کاب در اختیار قرار داده شده بود، تعیین ژنوتیپ شدند. برای هر نمونه ۵۴۳۴۰ نشان‌گر چندشکلی تک‌نوکلئوتیدی تعیین شد. این پرندگان نتاج حاصل از ۸ خانواده برادر – خواهر ناتنی بودند.

مراحل کنترل کیفیت بر روی داده‌های اصلی با استفاده از نرم افزار PLINK 1.9 انجام شد (Purcell *et al.*, 2007). نشان‌گرهای با نرخ فراخوانی کمتر از ۹۵ درصد یا حداقل فراوانی آللی (Minor allele frequency, MAF) کمتر از ۵ درصد، ژنوتیپ از دست رفته بیش از یک درصد و عدم تعادل هاردی واینبرگ در سطح احتمال کمتر از 10^{-6} حذف شدند. پس از فرآیند کنترل کیفیت، ۴۵۵۱۲ نشان‌گر و ۳۰۰ پرنده برای بررسی نهایی باقی ماندند. قبل از اجرای روش‌های یادگیری ماشین و رگرسیون خطی جهت شناسایی نشان‌گرهای مهم، داده های فنوتیپی مورد نظر برای اثرات جنس و نوبت جوجه‌کشی با روش تابعیت خطی تصحیح شدند.

روش‌های مورد استفاده جهت شناسایی نشان‌گرهای با اهمیت

مدل خطی

مدل خطی به صورت زیر است:

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{Z}\mathbf{q} + \mathbf{e}$$

که در آن، $\mathbf{y}$ بردار مقادیر فنوتیپی تصحیح شده برای وزن بدن در سنین ۶ یا ۹ هفتگی، $\mathbf{1}$ بردار n بعدی از یک‌ها، $\boldsymbol{\mu}$ میانگین جمعیت، $\mathbf{q}$ بردار نامعلوم اثر نشان‌گر، به عنوان تابعیت ثابت مشاهده از ژنوتیپ، $\mathbf{Z}$ بردار حاوی ژنوتیپ‌های نشان‌گر با مقادیر ۰ برای ژنوتیپ A_1A_1، ۱ برای ژنوتیپ A_1A_2 و ۲ برای ژنوتیپ A_2A_2 ، و $\mathbf{e}$ بردار مقادیر باقی‌مانده با فرض $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ است. فرض می‌شود که σ_e^2 واریانس باقیمانده و $\mathbf{I}$ ماتریس همانی است. آزمون ارتباط ژنتیکی با استفاده از دستور "--Linear" نرم افزار PLINK 1.9 انجام شد (Purcell *et al.*, 2007). در این مدل، ۱۰ چندشکلی تک‌نوکلئوتیدی برتر براساس مقادیر p-value از نتایج مدل خطی انتخاب شدند. لازم به توضیح است که دو عامل ساختار جمعیت و مولفه‌های اصلی، به این دلیل که قابل برازش در روش‌های یادگیری ماشین نبودند، در مدل فوق منظور نشدند.

جنگل تصادفی

روش جنگل‌تصادفی، یک الگوریتم ترکیبی از چندین درخت تصمیم است. در واقع، آن‌ها یک مجموعه‌ای از درختان تصمیم در یک جنگل ایجاد می‌کنند که می‌توانند پیش‌بینی بهتری نسبت به یک درخت فردی داشته باشند. در الگوریتم RF، برای رشد هر درخت از یک نمونه bootstrap (نمونه‌گیری با جایگزینی) از داده‌های آموزش اصلی استفاده می‌شود. داده‌های آموزشی شامل اطلاعات ژنوتیپی و فنوتیپ های تصحیح شده می‌باشد. برای تشکیل هر درخت براساس اطلاعات ژنوتیپی و فنوتیپی، یک نمونه با جایگزینی از داده‌های آموزشی گرفته و درخت روی نمونه گرفته شده رشد می‌کند و در هر گره از تعدادی از SNP ها که به صورت تصادفی انتخاب شده‌اند، یک SNP با توجه به قدرت تفکیک بهتر انتخاب می شود و گره براساس ژنوتیپ آن SNP منشعب و فنوتیپ‌ها بر اساس آن SNP به برگ یا گره بعدی تفکیک می‌شوند. در مراحل بعد و در گره های دیگر از نمونه‌های دیگر SNP، یک SNP انتخاب شده و انشعاب گره‌ها ادامه می‌یابد تا درخت کامل شود و همه‌ی فنوتیپ‌ها در برگ‌های درخت قرار بگیرند. به همین صورت درخت‌های بعدی شکل می‌گیرند. الگوریتم جنگل تصادفی، نتیجه را با میانگین‌گیری از خروجی‌های حاصل از همه درختان در جنگل پیش‌بینی می‌کند (Breiman, 2001). در هر بار bootstrap از اطلاعات، برخی اطلاعات هرگز نمونه‌گیری نمی‌شوند و برخی دیگر چند بار نمونه‌گیری می‌شوند. به عبارت دیگر، هر داده ورودی برای برخی درختان، داده‌های خارج از کیسه (out of Bag) خواهند بود

که در ایجاد برخی از درختان مشارکت ندارند. این داده‌ها به عنوان یک اعتبارسنج داخلی، که از طریق برآورد خطای خارج از کیسه (out of Bag error) انجام می‌شود، برای هر درخت عمل می‌کنند.

برای محاسبه اهمیت هر چندشکلی تک‌نوکلئوتیدی، از روش جای‌گذاری تصادفی (Permutation) استفاده می‌شود. ابتدا خطای پیش‌بینی با استفاده از نمونه‌های خارج از کیسه در درخت مربوطه محاسبه می‌شود. سپس مقادیر ژنوتیپی هر چندشکلی تک‌نوکلئوتیدی به صورت تصادفی جای‌گذاری می‌شوند و خطای پیش‌بینی نمونه‌های خارج از کیسه مجدداً محاسبه می‌شود. تفاوت بین خطای نمونه‌های خارج از کیسه برای چندشکلی تک‌نوکلئوتیدی با مقادیر ژنوتیپی اصلی و مقادیر ژنوتیپی جای‌گذاری شده به شکل تصادفی (میانگین در تمام درختان در جنگل) نشان‌دهنده اهمیت یا قابلیت پیش‌بینی آن چندشکلی تک‌نوکلئوتیدی خاص است که به صورت درصد افزایش در میانگین مربعات خطا (percentage of increase in MSE) مشخص می‌شود (Nicodemus *et al.*, 2010; Li *et al.*, 2018). هر چقدر این مقدار بیشتر باشد نشان می‌دهد نبود چندشکلی تک‌نوکلئوتیدی مربوطه در نمونه، باعث افزایش خطای پیش‌بینی شده و بنابراین از اهمیت بیشتری برخوردار است.

در مطالعه حاضر، ۳ پارامتر مهم روش RF برای تعیین مقدار مناسب آن‌ها در نظر گرفته شد که شامل تعداد درخت (Ntree) با مقادیر ۱۰۰۰ تا ۵۰۰۰ درخت (با فاصله ۱۰۰۰ درخت)، تعداد نمونه‌های انتخاب شده برای تشکیل هر گره (mtry) با مقادیر sqrt(p)، p/3، (p/3)×0.5 و (p/3)×2، که p نشان دهنده تعداد نشان‌گرها است و حداقل تعداد نشان‌گرها در گره‌های پایانی (node size) با مقادیر ۵، ۱۰ و ۱۵ بودند. از ریشه میانگین مربعات خطا (Root Mean Squared Error, RMSE) برای انتخاب بهترین ترکیب از پارامترها استفاده شد. برای اجرای روش RF از بسته نرم افزاری random Forest در نرم افزار R استفاده شد (Liaw and Wiener., 2015).

گرادیان بوستینگ

توابع اصلی استفاده شده در الگوریتم گرادیان بوستینگ (GB)، یادگیرنده‌های ضعیف مانند درخت‌های تصمیم (decision stump) هستند. در این روش سعی می‌شود تعدادی از یادگیرنده‌های ضعیف و کمکی تولید شود که به هم پیوسته بوده و با یک طرح آموزش مرحله‌ای برای کاهش خطا در پیش‌بینی کار کنند. هدف این الگوریتم، افزایش توان یک یادگیرنده قوی از مجموعه‌ای از یادگیرنده‌های ضعیف است. در این روش، یک یادگیرنده پایه به صورت ترتیبی به باقیمانده‌های درخت قبلی اضافه می‌شود و انتظار می‌رود با تمرکز بر روی داده‌های نادرست پیش‌بینی شده در درخت قبلی، نرخ خطا در درخت بعدی کاهش یابد و تا زمانی که نرخ خطا در حال کاهش باشد، الگوریتم بوستینگ ادامه خواهد داشت. در مطالعه حاضر، نشان‌گرهای مهم در روش GB با استفاده از تأثیر نسبی (relative influence) شناسایی می‌شوند که میانگین کاهش میانگین مربعات خطا (MSE) در همه درخت‌ها را محاسبه می‌کند زمانی که چندشکلی تک‌نوکلئوتیدی خاصی برای تقسیم داده‌ها استفاده می‌شود (Friedman, 2001).

سه پارامتر مهم روش GB برای تعیین مقدار مناسب آن‌ها، شامل تعداد درخت (Ntree) با مقادیر ۱۰۰۰ تا ۵۰۰۰ درخت (با فاصله ۱۰۰۰ درخت)، حداقل تعداد نشان‌گرها در گره‌های پایانی (n.minobinnode) با مقادیر ۵، ۱۰ و ۱۵ و ضریب اختصاص داده شده به درختان (shrinkage) با مقادیر ۰/۱، ۰/۰۷، ۰/۰۴ و ۰/۰۱ بودند. از بسته نرم افزاری gbm در نرم افزار R برای اجرای روش GB استفاده شد (Ridgeway *et al.*, 2013). برای روش‌های RF و GB تنظیمات پارامترها از طریق جستجوی شبکه با استفاده از 3-fold cross-validation بر روی ۷۵ درصد نمونه‌های تصادفی انجام شد.

غربالگری گستره ژنوم برای نشان‌گرهای برتر

تمام نشان‌گرها بر اساس اهمیت آن‌ها پس از اجرای مدل خطی و روش‌های یادگیری ماشین بر اساس بهترین ترکیب پارامترها رتبه‌بندی شدند. پس از رتبه‌بندی، ده نشان‌گر برتر از هر روش برای صفات وزن بدن مشخص شد. علاوه بر این، ژن‌های قرار گرفته

در فاصله Mb−۱ بالا و پایین از ۳ نشان‌گر برتر، که از هر روش در ناحیه ژنومی تشخیص داده شده بود، با استفاده از پایگاه داده‌های NCBI و Ensemble از ژنوم مرجع مرغ (Gallus-Gallus) مشخص شدند (Yates *et al*., 2016). جهت کاوش فرآیندهای بیولوژیکی و عملکردی ژن‌ها، هستی‌شناسی ژن با استفاده از نرم افزار آنلاین DAVID (http://david.abcc.ncifcrf.gov/) استفاده شد.

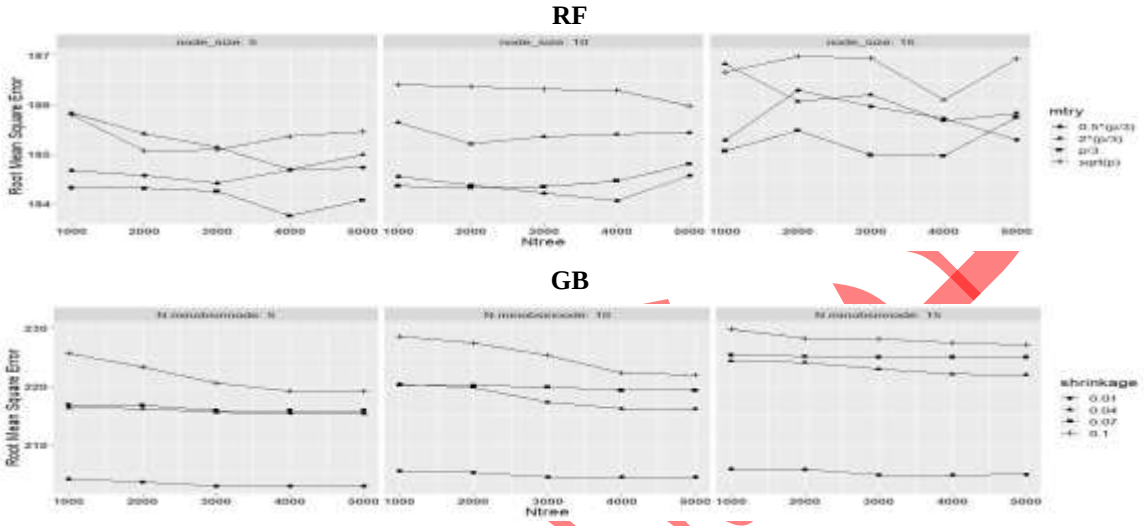
یافته های پژوهش و بحث

نتایج ترکیب مختلف از پارامترهای مهم روش‌های یادگیری ماشین براساس ریشه میانگین مربعات خطا برای صفات وزن بدن در ۶ و ۹ هفتگی به ترتیب در شکل‌های ۱ و ۲ نشان داده شده است. در روش جنگل تصادفی، ترکیب تعداد نمونه‌های انتخاب شده برای تشکیل هر گره، تعداد درخت و حداقل تعداد نشان‌گرها در گره‌های پایانی به ترتیب برابر p/3، ۴۰۰۰ و ۵ برای صفت وزن بدن در ۶ هفتگی و 0.5×p/3، ۳۰۰۰ و ۵ برای صفت وزن بدن در ۹ هفتگی حداقل خطا را ایجاد کرد. در روش گرادیان بوستینگ، ترکیب تعداد درخت، حداقل تعداد نشان‌گرها در گره‌های پایانی و ضریب اختصاص داده شده به درختان به ترتیب برابر ۳۰۰۰، ۵ و ۰/۰۱ برای هر دو صفت وزن بدن انتخاب شدند. در مطالعات پویش ژنومی پیشنهاد شده است که مقدار پارامتر تعداد متغیر انتخاب شده در هر گره درخت (mtry) برای روش جنگل تصادفی باید از ۰/۱ تعداد SNPها بیشتر باشد (Goldstein *et al*., 2010) که در مطالعه حاضر برای این پارامتر، به ترتیب، ۱۵۱۷۱ و ۷۵۸۵ SNP برای صفت وزن ۶ و ۹ هفتگی در نظر گرفته شد. مقدار مناسب حداقل اندازه‌ی گره‌های پایانی در هر دو روش (node size در جنگل تصادفی و n.minobinnode در گرادیان بوستینگ) برای هر دو صفت برابر ۵ تعیین شد. هر اندازه مقدار این پارامتر بزرگتر باشد گره و برگ کم‌تری تشکیل خواهد شد و در نتیجه درختان کوچک‌تری تولید می‌شوند که ممکن است خیلی از نشان‌گرها در تشکیل درخت شرکت نکنند و در نتیجه کارایی مدل برای انتخاب نشان‌گرهای با اهمیت کاهش پیدا کند، بنابراین پیشنهاد شده است از مقادیر کوچک برای این پارامتر استفاده شود (Boulesteix *et al*., 2012). در هر دو روش با افزایش تعداد درخت، احتمال اینکه بیشتر نشان‌گرهای نمونه‌گیری و اهمیت آن‌ها سنجیده شود، افزایش می‌یابد. به عبارت دیگر، هر نشان‌گر فرصت آن را خواهد داشت که حداقل یک بار نمونه‌گیری شود. بنابراین، زمانی که تعداد نشان‌گرها زیاد است نباید از مقادیر کوچک یا پیش فرض برنامه برای تعداد درخت استفاده شود (Boulesteix *et al*., 2012). در روش گرادیان بوستینگ، ضریب ۰/۱۰ اختصاص داده شده به درختان برای هر دو صفت، کمترین خطا را ایجاد کرد. این پارامتر، میزان مشارکت هر درخت در مدل را نشان می دهد و مقدار آن معمولا با تعداد درخت رابطه عکس دارد، به طوری که با افزایش این پارامتر می‌توان تعداد درخت مورد نیاز جهت ایجاد مدل را کاهش داد. این پارامتر با مقداری کوچک‌تر از ۰/۱ و تعداد درخت زیاد جهت تحلیل دقیق‌تر در روش بوستینگ پیشنهاد شده است (Friedman, 2002).

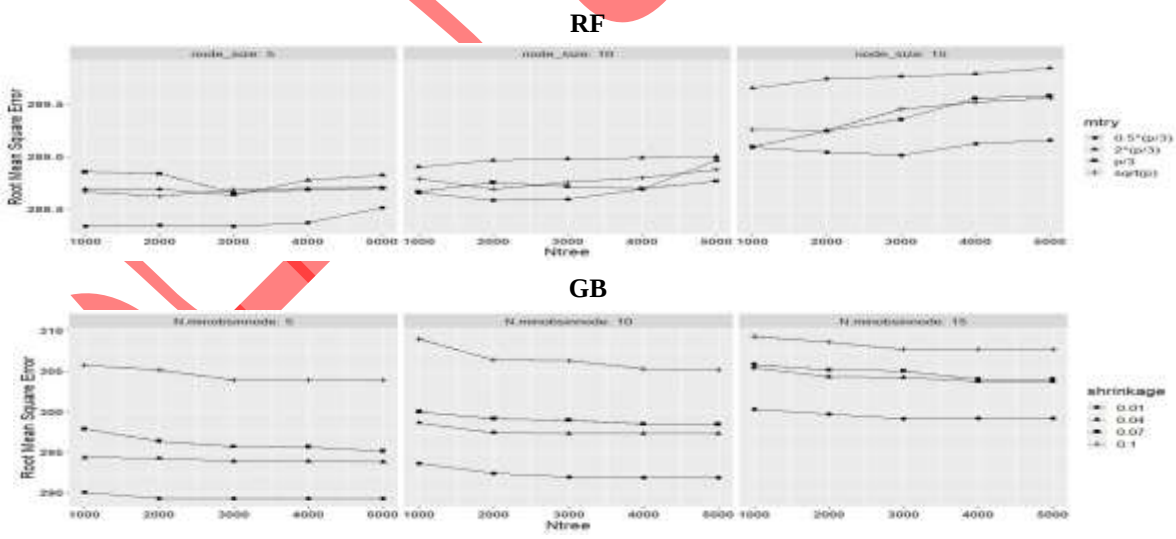
پلات منهتن برای صفات وزن بدن در ۶ هفتگی و ۹ هفتگی با استفاده از روش‌های مدل خطی، جنگل تصادفی و گرادیان بوستینگ، به ترتیب در شکل‌های ۳ و ۴ نشان داده شده است. در روش جنگل تصادفی، اهمیت نشانگرها به صورت منفی، صفر و مثبت نشان داده می شود در حالی که در روش گرادیان بوستینگ اهمیت نشانگرها، صفر یا مثبت است (Li *et al*., 2018). برای صفت وزن بدن در ۶ و ۹ هفتگی به ترتیب ۴۷ درصد و ۴۸ درصد از نشانگرها در جنگل تصادفی و ۲۶ درصد و ۱۶ درصد از نشانگرها در روش گرادیان بوستینگ دارای اهمیت مثبت بودند. با استفاده از دو روش یادگیری ماشین در کروموزوم‌های ۴ و ۷ برای صفت وزن بدن در ۶ هفتگی و در کروموزوم‌های ۶ و ۸ برای صفت وزن بدن در ۹ هفتگی، نشانگرهای با اهمیتی کشف شدند.

تعداد ده نشان‌گر برتر شناسایی شده از هر سه روش برای صفات وزن بدن در ۶ و ۹ هفتگی به ترتیب در جداول ۲ و ۳ نشان داده شده است. در مدل خطی، دامنه مقدار p-value برای ده نشان‌گر برتر از $۲/۰۱ \times ۱۰^{-۵}$ تا $۱/۵۳ \times ۱۰^{-۴}$ برای وزن ۶ هفتگی و از $۲/۵۱ \times ۱۰^{-۵}$ تا $۱/۳۱ \times ۱۰^{-۴}$ برای وزن ۹ هفتگی بود. براساس این مقادیر، نشان‌گرهای GGaluGA141221 و GGaluGA142838 به ترتیب برای وزن ۶ و ۹ هفتگی مهم‌ترین نشان‌گرها بودند. در روش جنگل تصادفی، دامنه مقدار درصد افزایش خطای میانگین مربعات برای ده نشان‌گر برتر از ۰/۴۱ تا ۰/۹۶ برای وزن ۶ هفتگی و از ۰/۳۱ تا ۰/۶۶ برای وزن ۹ هفتگی

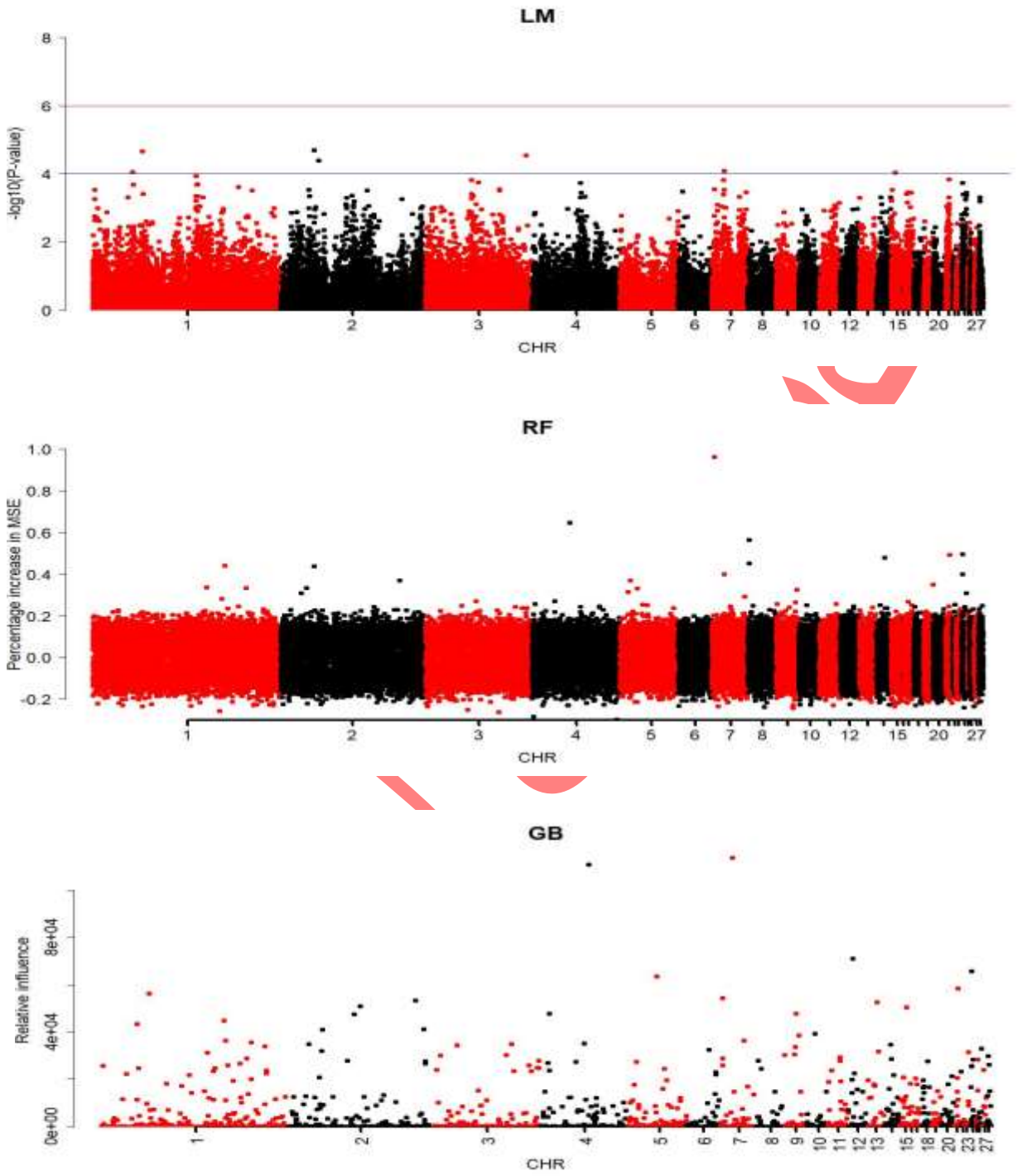
بود. در این روش، برای وزن بدن در ۶ و ۹ هفتگی، مهم‌ترین نشان‌گرها، به ترتیب، GGaluGA308573 و Gga_rs15763229 بودند. دامنه مقدار تاثیر نسبی در روش گرادیان بوستینگ برای وزن ۶ هفتگی از ۵۲۶۲۳/۳۵ تا ۱۱۳۹۲۹/۴ و برای وزن ۹ هفتگی از ۴۰۴۷۹ تا ۱۶۳۳۵۲ بود. با در نظر گرفتن این مقادیر، دو نشان‌گر Gga_rs13743072 و Gga_rs13614212، به ترتیب، مهم‌ترین نشان‌گرهای کشف شده برای وزن‌های ۶ و ۹ هفتگی بودند.



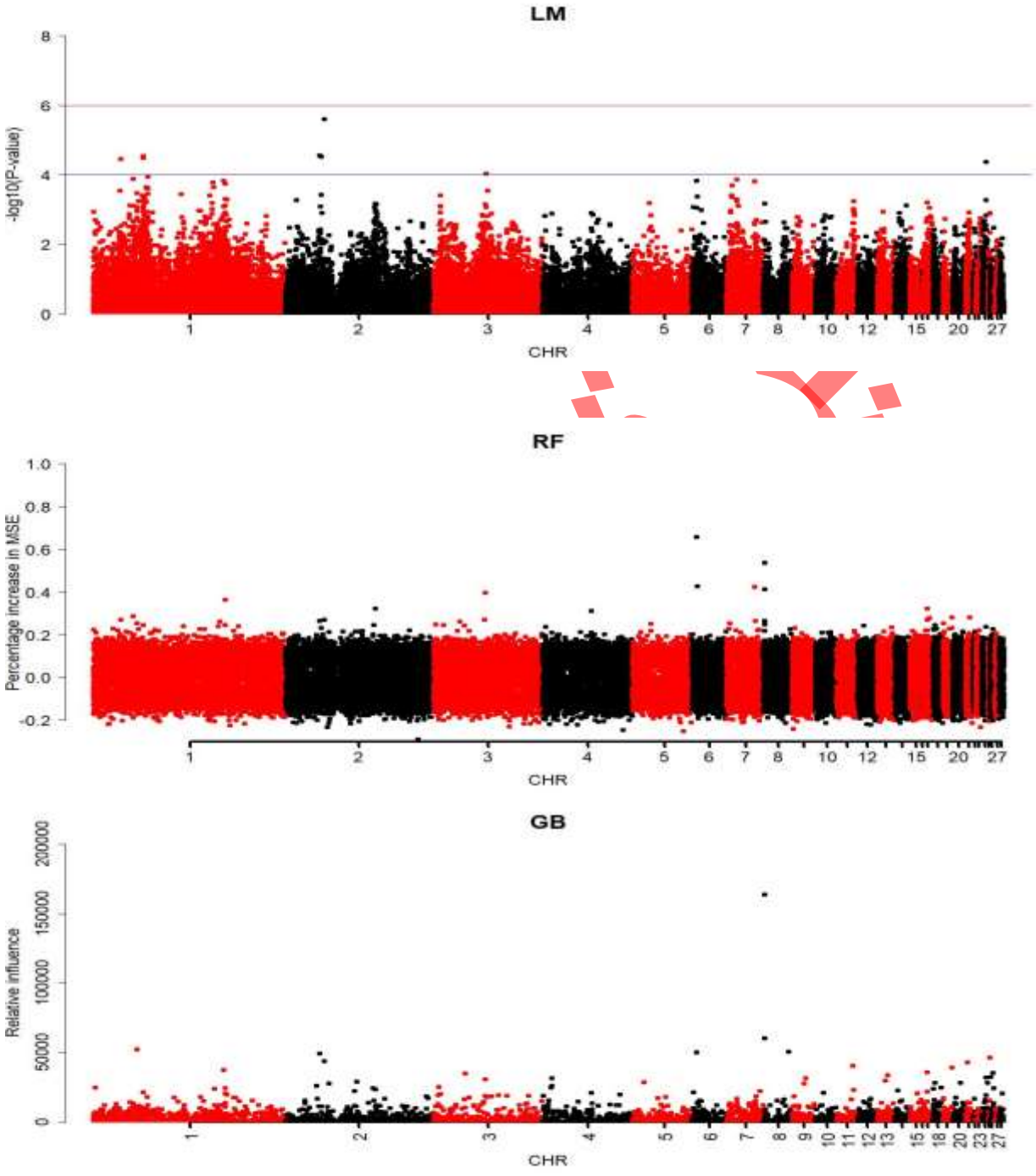
شکل ۱. ارتباط بین ترکیب‌های مختلف پارامترها و ریشه میانگین مربعات خطا برای روش‌های جنگل تصادفی (RF) و گرادیان بوستینگ (GB) در صفت وزن بدن در ۶ هفتگی. mtry: تعداد نشان‌گرهای انتخاب شده برای ساخت درخت، P: تعداد کل SNP ها، محور X به اندازه درخت جنگل (NTree) است.



شکل ۲. ارتباط بین ترکیب‌های مختلف پارامترها و ریشه میانگین مربعات خطا برای روش‌های جنگل تصادفی (RF) و گرادیان بوستینگ (GB) در صفت وزن بدن در ۹ هفتگی. mtry: تعداد نشان‌گرهای انتخاب شده برای ساخت درخت، P: تعداد کل SNP ها، محور X به اندازه درخت جنگل (NTree) است.



شکل۳. پلات منهتن برای صفت وزن بدن در ۶ هفتگی با استفاده از روش‌های مدل خطی (LM)، جنگل تصادفی (RF) و گرادیان بوستینگ (GB)



شکل٤. پلات منهتن برای صفت وزن بدن در ۹ هفتگی با استفاده از روش‌های مدل خطی (LM)، جنگل تصادفی (RF) و گرادیان بوستینگ (GB)

جدول ۲. ده نشان‌گرهای برتر از مدل خطی (LM)، جنگل‌های تصادفی (RF) و گرادیان بوستینگ (GB) برای صفت وزن بدن در ۶ هفتگی

Method	Rank	Chr.	SNP	Position (bp)	p-value	MAF
LM	۱	۲	GGaluGA141221	۳۵۹۱۹۵۲۹	۲/۰۱E−۰۵	۰/۱۷
	۲	۱	Gga_rs13866016	۵۱۴۲۵۹۷۸	۲/۱۸E−۰۵	۰/۴۰
	۳	۳	GGaluGA237887	۱۰۷۵۰۸۲۱۹	۲/۸۶E−۰۵	۰/۱۴
	۴	۲	GGaluGA142838	۴۰۵۲۸۶۶۶	۴/۰۳E−۰۵	۰/۳۳
	۵	۷	Gga_rs13743072	۱۱۹۳۲۶۸۰	۸/۲۱E−۰۵	۰/۳۴
	۶	۱	GGaluGA014205	۴۲۴۹۸۳۳۲	۸/۹۲E−۰۵	۰/۰۷
	۷	۱۵	Gga_rs13629367	۴۳۲۰۰۲۰	۹/۳۹E−۰۵	۰/۰۶
	۸	۱	Gga_rs13918443	۱۱۰۶۵۴۵۱۲	۰/۰۰۰۱۱۸۷	۰/۱۹
	۹	۲۱	Gga_rs14284004	۳۵۰۲۵۶۳	۰/۰۰۰۱۴۹۵	۰/۱۶
	۱۰	۷	Gga_rs14607739	۱۲۷۳۲۸۴۶	۰/۰۰۰۱۵۲۹	۰/۱۶
Method	**Rank**	**Chr.**	**SNP**	**Position (bp)**	**Percentage increase in MSE**	
RF	۱	۷	GGaluGA308573	۲۵۶۴۱۳۲	۰/۹۶	۰/۲۸
	۲	۴	GGaluGA255033	۳۹۰۳۷۸۶۸	۰/۶۵	۰/۰۷
	۳	۸	Gga_rs13614212	۱۳۹۴۳۶۹	۰/۵۶	۰/۲۳
	۴	۲۴	Gga_rs14292823	۱۱۵۶۸۴۰	۰/۵۰	۰/۴۱
	۵	۲۱	GGaluGA184364	۴۳۰۹۲۶۳	۰/۴۹	۰/۲۶
	۶	۱۴	Gga_rs14076673	۸۷۹۸۲۷۸	۰/۴۹	۰/۰۹
	۷	۸	GGaluGA322130	۱۵۴۱۱۵۳	۰/۴۸	۰/۲۱
	۸	۱	Gga_rs10727436	۱۴۱۰۸۱۷۸۸	۰/۴۵	۰/۲۹
	۹	۲	GGaluGA141221	۳۵۹۱۹۵۲۹	۰/۴۴	۰/۱۷
	۱۰	۲۴	Gga_rs14292795	۱۱۳۰۳۷۶	۰/۴۱	۰/۴۱
Method	**Rank**	**Chr.**	**SNP**	**Position (bp)**	**Relative influence**	
GB	۱	۷	Gga_rs13743072	۱۱۹۳۲۶۸۰	۱۱۳۹۲۹	۰/۳۴
	۲	۴	GGaluGA258772	۴۹۸۶۳۷۶۱	۱۱۱۰۰۵	۰/۱۸
	۳	۱۲	Gga_rs14034395	۴۹۲۴۱۰۹	۷۰۹۹۳	۰/۳۶
	۴	۲۴	Gga_rs15214346	۱۸۹۰۱۴۰	۶۵۷۳۹	۰/۱۵
	۵	۵	Gga_rs14527434	۳۰۱۸۰۳۹۳	۶۳۶۵۶	۰/۱۲
	۶	۲۱	GGaluGA184364	۴۳۰۹۲۶۳	۵۸۵۵۶	۰/۲۶
	۷	۱	GGaluGA016965	۵۱۴۸۸۷۵۲	۵۶۳۴۵	۰/۲۸
	۸	۷	GGaluGA308573	۲۵۶۴۱۳۲	۵۴۳۶۸	۰/۲۹
	۹	۲	Gga_rs14251348	۱۳۴۹۰۹۲۹۸	۵۳۳۰۴	۰/۴۷
	۱۰	۱۳	GGaluGA094688	۱۰۸۲۸۸۶۵	۵۲۶۲۳	۰/۳۰

Chr.: شماره کروموزوم، Rank: رتبه نشان‌گر، SNP: نام نشان‌گر، MAF: فراوانی آلل کمیاب.

جدول ۳. ده نشان‌گرهای برتر از مدل خطی(LM) ، جنگل‌های تصادفی (RF) و گرادیان بوستینگ (GB) برای صفت وزن بدن در 9 هفتگی

Method	Rank	Chr.	SNP	Position (bp)	p-value	MAF
LM	۱	۲	GGaluGA142838	۴۱۰۹۳۸۶۸	۲/۵۱E−۰۵	۰/۳۳
	۲	۲	GGaluGA141221	۳۵۹۱۹۵۲۹	۲/۷۹E−۰۵	۰/۱۷
	۳	۱	Gga_rs13865536	۵۰۶۸۵۶۲۲	۲/۹۱E−۰۵	۰/۴۶
	۴	۲	GGaluGA141644	۳۷۴۱۵۱۸۲	۳/۰۳E−۰۵	۰/۲۸
	۵	۱	GGaluGA017356	۵۲۳۵۸۹۷۰	۳/۲۱E−۰۵	۰/۴۳
	۶	۱	Gga_rs13747634	۲۸۴۹۷۱۹۲	۳/۵۴E−۰۵	۰/۳۰
	۷	۲۴	Gga_rs14295712	۳۹۱۴۶۹۹	۴/۲۱E−۰۵	۰/۲۹
	۸	۳	Gga_rs16277926	۵۴۳۳۰۷۰۶	۹/۲۴E−۰۵	۰/۴۵
	۹	۱	Gga_rs13654297	۵۷۳۵۵۲۰۶	۰/۰۰۰۱۱۲	۰/۲۵
	۱۰	۱	Gga_rs14814275	۴۱۴۴۷۹۳۷	۰/۰۰۰۱۳۱	۰/۴۲
Method	**Rank**	**Chr.**	**SNP**	**Position (bp)**	**Percentage increase in MSE**	
RF	۱	۶	Gga_rs15763229	۶۰۵۹۴۱۰	۰/۶۶	۰/۰۸
	۲	۸	Gga_rs13614212	۱۳۹۴۳۶۹	۰/۵۴	۰/۲۳
	۳	۶	GGaluGA295472	۶۳۰۹۳۶۳	۰/۴۲	۰/۰۶
	۴	۷	GGaluGA318133	۲۹۲۵۵۶۱۶	۰/۴۲	۰/۴۴
	۵	۸	GGaluGA322130	۱۵۴۱۱۵۳	۰/۴۱	۰/۲۱
	۶	۳	Gga_rs14359646	۵۴۰۷۶۶۸۹	۰/۳۹	۰/۳۱
	۷	۱	Gga_rs13625427	۱۳۸۰۴۴۸۴۷	۰/۳۶	۰/۲۹
	۸	۲	GGaluGA158217	۹۳۸۲۵۶۲۱	۰/۳۲	۰/۴۶
	۹	۱۷	GGaluGA114401	۵۰۰۶۷۲۹	۰/۳۲	۰/۱۷
	۱۰	۴	GGaluGA258464	۵۱۲۲۶۹۵۱	۰/۳۱	۰/۲۴
Method	**Rank**	**Chr.**	**SNP**	**Position (bp)**	**Relative influence**	
GB	۱	۸	Gga_rs13614212	۱۳۹۴۳۶۹	۱۶۳۵۶۲	۰/۲۳
	۲	۸	GGaluGA322130	۱۵۴۱۱۵۳	۶۰۱۱۲	۰/۲۱
	۳	۱	Gga_rs13858398	۴۳۹۵۹۹۰۲	۵۲۲۹۱	۰/۳۰
	۴	۸	Gga_rs14654881	۲۶۳۶۶۷۷۵	۵۰۵۱۶	۰/۲۳
	۵	۶	Gga_rs15763229	۶۰۵۹۴۱۰	۵۰۱۴۶	۰/۰۸
	۶	۲	GGaluGA141221	۳۵۹۱۹۵۲۹	۴۹۴۲۴	۰/۱۷
	۷	۲۵	Gga_rs16723884	۱۹۴۴۰۹۶	۴۶۱۲۵	۰/۱۶
	۸	۲	GGaluGA142838	۴۱۰۹۳۸۶۸	۴۳۵۶۲	۰/۳۳
	۹	۲۱	Gga_rs16178040	۱۶۱۵۸۴۶	۴۳۰۷۵	۰/۰۸
	۱۰	۱۱	Gga_rs14695411	۱۶۶۹۸۶۴۴	۴۰۴۷۹	۰/۲۸

Chr.: شماره کروموزوم، Rank: رتبه نشان‌گر، SNP: نام نشان‌گر، MAF: فراوانی آلل مینور.

در روش‌های مورد استفاده در مطالعه حاضر، نشان‌گرهای متفاوتی از لحاظ اهمیت رتبه‌بندی شدند ولی برخی از نشان‌گرها حداقل در دو روش، جزء ده نشان‌گرهای برتر بودند. در وزن ۶ هفتگی، نشان‌گرهای GGaluGA308573 و GGaluGA184364 در دو روش جنگل تصادفی و گرادیان بوستینگ، نشان‌گر GGaluGA141221 در دو روش جنگل تصادفی و مدل خطی، و نشان‌گر Gga_rs13743072 در دو روش مدل خطی و گرادیان بوستینگ با رتبه‌های متفاوت، مشترک و جز ده نشان‌گر برتر بودند. در وزن ۹ هفتگی، سه نشان‌گر Gga_rs15763229، Gga_rs13614212 و GGaluGA322130 در بین دو روش جنگل تصادفی و گرادیان

بوستینگ و نشان‌گر GGaluGA142838 در بین دو روش مدل خطی و گرادیان بوستینگ با رتبه‌های متفاوت جزو ده نشان‌گر برتر بودند. در هردو صفت، نشان‌گرهای مشترک بین دو روش یادگیری ماشین بیشتر بود که می‌تواند اهمیت آن‌ها بر صفات مورد نظر را نشان دهد. ممکن است در مدل‌های خطی به دلیل فرضیات و سخت‌گیری‌های موجود در معنی‌دار شدن نشان‌گرها، امکان شناسایی آن‌ها میسر نشده باشد.

ژن‌های مرتبط با ۳ نشان‌گر مهم هر روش برای صفات وزن بدن در ۶ و ۹ هفتگی در جدول ۴ نشان داده شده است. نشان‌گر GGaluGA141221 توسط مدل خطی در دو صفت وزن بدن جز ۳ نشان‌گر برتر و برای وزن بدن در ۶ و ۹ هفتگی، به ترتیب، دارای رتبه ۱ و ۲ بود. این نشان‌گر، که توسط الگوریتم‌های یادگیری ماشین هم جز نشان‌گرهای برتر و با اهمیت برای هر دو صفت وزن بدن شناخته شده است، بر روی کروموزوم ۲ قرار دارد و با ژن *SGOL1* در ارتباط است. با وجود این‌که در مطالعه حاضر، داده‌ها برای ساختار جمعیت تصحیح نشدند، اما برخی نتایج حاصل از آن مثل شناسایی نشان‌گر مرتبط با ژن *SGOL1*، با یافته‌های Emrani و همکاران (۲۰۱۷) مطابقت داشت. وجود تفاوت در سایر نشان‌گرهای مطالعه حاضر با یافته‌های Emrani و همکاران (۲۰۱۷)، علی‌رغم استفاده از داده‌های مشابه، به این دلیل بود که این محققین پویش ژنومی داده‌ها را با مدل خطی انجام دادند که در آن اثرات ساختار جمعیت، وزن بدن جوجه‌های یک‌روزه و پلی ژنیک منظور شده بود. حذف این عوامل در مدل خطی مطالعه حاضر به این دلیل بود که نتایج حاصل از این مدل با نتایج دو روش یادگیری ماشین (جنگل تصادفی و گرادیان بوستینگ)، که امکان برازش این عوامل در آن روش‌ها وجود نداشت (Arabnejad *et al.*, 2020)، قابل مقایسه باشد.

نشان‌گر Gga_rs13614212 یکی دیگر از نشان‌گرهای مهم برای هر دو صفت وزن بود که توسط الگوریتم‌های یادگیری ماشین شناسایی شد. این نشان‌گر، که در وزن ۶ هفتگی توسط روش جنگل تصادفی و در وزن ۹ هفتگی توسط هر دو روش یادگیری ماشین نیز به‌عنوان ۳ نشان‌گر برتر شناسایی شده بود، بر روی کروموزوم ۸ و در درون ژن *CAMSAP2* قرار داشت. پروتئین حاصل از این ژن به سر آنتی‌سانتروزومال میکروتوبول‌ها متصل می‌شود و این امکان را به آن می‌دهد تا دینامیک، ساختار و پلیمریزاسیون میکروتوبول‌ها را کنترل کند (Jiang *et al.*, 2014). این ژن از جمله ژن‌های معنی‌دار برای صفات لاشه در گاوهای گوشتی (Srikanth *et al.*, 2020) و ضریب تبدیل خوراک در جوجه های گوشتی (Wang *et al.*, 2019) گزارش شده است.

در صفت وزن بدن در ۶ هفتگی، دو نشان‌گر GGaluGA308573 و Gga_rs13743072 در هر ۳ روش اهمیت زیادی داشتند. نشان‌گر GGaluGA308573 در روش‌های مدل خطی، جنگل تصادفی و گرادیان بوستینگ به ترتیب رتبه ۱۸، ۱ و ۸ و نشان‌گر Gga_rs13743072 به ترتیب رتبه ۵، ۱۱ و ۱ را داشتند. نشان‌گر GGaluGA308573 بر روی کروموزوم ۷ و در درون ژن *MAP2* قرار داشت. این ژن یک بخش اصلی از سیستم عصبی است و بیشتر در دندریت‌ها یافت شده و به عنوان فعال‌کننده و استحکام دهنده میکروتوبول و کنترل‌کننده سیستم‌های میکروتوبول در دندریت‌های نورون‌ها شناخته شده است که منجر به کشش دندریت می‌شود (Harada *et al.*, 2002). در مطالعه Fehm و همکاران (۲۰۰۴) گزارش شده است که وزن بدن در انسان از طریق سیستم عصبی مرکزی تنظیم می‌شود، زیرا گیرنده‌های گلوکوکورتیکوئید و مینرالوکورتیکوئید در نورون‌ها، تعیین کننده تعادل بین فرآیندهای تخصیص قند و مصرف غذا هستند. ژن *MAP2* به عنوان یکی از ژن‌های مؤثر در تمایز بین جمعیت گوسفند غیر تجاری در آفریقای جنوبی (Nguni sheep)، که برای ویژگی تولیدی خاصی انتخاب نشده، و نژادهای تجاری (Mutton Merino and Dorset Horn)، که برای رشد و تولید گوشت انتخاب شده‌اند، معرفی شد (Dzomba *et al.*, 2020). نشان‌گر Gga_rs13743072 در کروموزوم ۷ و در بالا دست ژن *FAM117B* قرار داشت. در مطالعات گزارش شده است که این ژن در انسان بر سطوح کلسترول بدن (Willer *et al.*, 2013) و در جوجه‌های گوشتی، بر توسعه و رشد جنین موثر است (Kanakachari *et al.*, 2021). نوع دیگری از خانواده این ژن (*FAM184B*) با همان توالی، به عنوان ژن مؤثر بر وزن بدن در سنین ۴، ۶، ۱۰ و ۱۲ هفته در جوجه‌های گوشتی گزارش شده است (Jin *et al.*, 2015).

جدول ٤. سه نشان‌گر مهم شناسایی شده با روش‌های مدل خطی، جنگل تصادفی و گرادیان بوستینگ و ژن‌های مرتبط با آن‌ها برای صفات وزن بدن در ۶ هفتگی (BW6) و ۹ هفتگی (BW9)

SNP	Chr.	Position (bp)	Rank in LM	Rank in RF	Rank in GB	P-value	PIncMSE	RI	Gene	Location
BWT6										
GGaluGA141221	۲	۳۵۹۱۹۵۲۹	۱	۹	۱۹	۲/۰۱E−۰۵	۰/۴۴	۴۰۸۹۳	SGOL1	Intron
Gga_rs13866016	۱	۵۱۴۲۵۹۷۸	۲	۹۰	۵۱۹	۲/۱۸E−۰۵	۰/۲۲	۱۱۲۲	CYTH4	Intron
GGaluGA237887	۳	۱۰۷۵۰۸۲۱۹	۳	۱۱۴	۷۰۴	۲/۸۶E−۰۵	۰/۲۱	۳۴۴	DTNB	Intron
GGaluGA308573	۷	۲۵۶۴۱۳۲	۱۸	۱	۸	۰/۰۰۰۳	۰/۹۶	۵۴۳۶۸	MAP2	Intron
GGaluGA255033	۴	۳۹۰۳۷۸۶۸	۲۴۳۷	۲	۹۸۱	۰/۰۳۵	۰/۶۵	۶۵	ACSL1	Intron
Gga_rs13614212	۸	۱۳۹۴۳۶۹	۵۰۰۰	۳	۹۴۳	۰/۰۸	۰/۵۶	۸۳	CAMSAP2	Intron
Gga_rs13743072	۷	۱۱۹۳۲۶۸۰	۵	۱۱	۱	۸/۲۱E−۰۵	۰/۴۰	۱۱۳۹۲۹	FAM117B	Upstream
GGaluGA258772	۴	۴۹۸۶۳۷۶۱	۱۷۷	۶۵۵	۲	۰/۰۰۱	۰/۱۷	۱۱۱۰۰۵	SLC4A4	Intron
Gga_rs14034395	۱۲	۴۹۲۴۱۰۹	۴۸۹۱	۹۹۴	۳	۰/۰۷	۰/۱۶	۷۰۹۹۳	TIMP4	Intron
BWT9										
GGaluGA142838	۲	۴۱۰۹۳۸۶۸	۱	۱۶	۸	۲/۵۱E−۰۵	۰/۲۸	۴۳۵۶۳	CNOT10	Intron
GGaluGA141221	۲	۳۵۹۱۹۵۲۹	۲	۲۲	۶	۲/۷۹E−۰۵	۰/۲۶	۴۹۴۲۴	SGOL1	Intron
Gga_rs13865536	۱	۵۰۶۸۵۶۲۲	۳	۹۸۵	۷۷۱	۲/۹۱E−۰۵	۰/۱۶	۵۱۷۰	MGAT3	Intron
Gga_rs15763229	۶	۶۰۵۹۴۱۰	۱۳	۱	۵	۰/۰۰۰۱	۰/۶۶	۵۰۱۴۶	-	Intergenic
Gga_rs13614212	۸	۱۳۹۴۳۶۹	۵۱	۲	۱	۰/۰۰۰۷	۰/۵۴	۱۶۳۵۶۲	CAMSAP2	Intron
GGaluGA295472	۶	۶۳۰۹۳۶۳	۳۵	۳	۱۲۸	۰/۰۰۰۴	۰/۴۲	۱۳۷۳۳	LncRNA	Intron
Gga_rs13614212	۸	۱۳۹۴۳۶۹	۵۱	۲	۱	۰/۰۰۰۷	۰/۵۴	۱۶۳۵۶۲	CAMSAP2	Intron
GGaluGA322130	۸	۱۵۴۱۱۵۳	۲۱۷۲	۵	۲	۰/۰۳	۰/۴۱	۶۰۱۱۲	-	Intergenic
Gga_rs13858398	۱	۴۳۹۵۹۹۰۲	۱۳۴	۶۸۴	۳	۰/۰۰۱	۰/۱۷	۵۲۲۹۱	LncRNA	Upstream

Chr.: تعداد کروموزوم، Position: موقعیت نشان‌گر، LM: مدل خطی، RF: جنگل تصادفی، GB: گرادیان بوستینگ، PIncMSE: درصد افزایش میانگین مربعات خطا، RI: تاثیر نسبی.

برای صفت وزن ۶ هفتگی، دو ژن *CYTH4* و *DTNB* در روش مدل خطی، ژن *ACSL1* در روش جنگل تصادفی و دو ژن *SLC4A4* و *TIMP4* در روش گرادیان بوستینگ در ارتباط با نشان‌گرهای با اهمیت شناسایی شدند. ژن *CYTH4*، ژن کدگذار برای تولید پروتینی به نام Cytohesin4 است که عضوی از خانواده‌ی پروتین‌های Cytohesine می‌باشد و در فعال‌سازی سیگنال‌های داخلی سلولی نقش دارد. این گروه پروتینی ممکن است در عملکردهای متعددی از جمله تنظیم سوخت وساز، تنظیم رشد و اندام سلولی و فعالیت‌های نورونی نقش داشته باشند (Teuliere *et al*., 2014; Cremonesi *et al*., 2012). ژن *DTNB*، که توسط Emrani و همکاران (۲۰۱۷) برای صفات رشد درجوجه‌های گوشتی گزارش شده است، در تولید کمپلکس پروتئین مرتبط با دیستروفین (DPC) نقش دارد. اختلال در این کمپلکس پروتئینی، که در سارکولم مشاهده می‌شود، با اشکال مختلف دیستروفی عضلانی همراه است. ژن *ACSL1* تولید پروتینی به همین نام در جوجه‌های گوشتی را کد می‌کند و نقش مهمی در متابولیسم چربی‌ها و اسیدهای چرب زنجیره بلند دارد. در جوجه‌های گوشتی، استفاده از ژن *ACSL1* می‌تواند به عنوان یکی از راهکارهای بهبود ترکیبات چربی در بافت عضلانی و سیستم ایمنی جوجه‌ها مورد استفاده قرار گیرد. این ژن می‌تواند بر رشد و توسعه عضلات جوجه‌های گوشتی تأثیر داشته و باعث بهبود کیفیت گوشت جوجه‌ها شود (Liu *et al* 2019., Tian *et al*., 2021). ژن *SLC4A4* پروتینی به نام NHE4 را کد می‌کند که نقش اصلی این پروتین در تنظیم غلظت یون هیدروژن و بی‌کربنات در سلول‌ها است. در جوجه‌های گوشتی، ژن

SLC4A4 نقش مهمی در تنظیم تعادل اسیدی–بازی و موازنه آب و الکترولیت دارد. این ژن در بافت‌های مخاطی روده‌ها و کلیه‌ها فعالیت می‌کند و در حفظ تعادل pH خون بدن نقش دارد. ژن *SLC4A4* از جمله ژن‌های تاثیرگذار در کنترل استرس گرمایی، چربی-های شکمی و رشد در پرندگان است (Resnyk *et al*., 2017; Bahadoran *et al*., 2018). ژن *TIMP4* کد کننده پروتین و عضوی از خانواده تیمپس (TIMPs) است که به عنوان مهارکننده طبیعی آنزیم‌های فعالیت متالوپروتئیناز (Metalloproteinases) عمل می‌کند. آنزیم‌های متالوپروتئیناز انواعی از آنزیم‌ها هستند که قابلیت تخریب و تفکیک پروتین‌ها را دارند. این آنزیم‌ها در فرآیندهایی مانند رشد و توسعه سلولی، ترمیم بافت‌های ملتهب و آسیب‌دیده نقش مهمی داشته و از جمله ژن‌های تاثیر گذار در رشد عضله در انسان و توسعه بافت روده جوجه‌های در حال رشد است (Brew *et al*., 2000; Hrabia *et al*., 2022). در مطالعه Yue و همکاران (۲۰۲۲) گزارش شده است که این ژن از جمله ژن‌های تاثیرگذار بر استحکام استخوان در پرندگان تخم‌گذار است.

برای صفت وزن بدن در ۹ هفتگی، نشان‌گرهای Gga_rs15763229 و GGaluGA322130، که به ترتیب بر روی کروموزوم های ۶ و ۸ قرار داشتند، توسط دو روش یادگیری ماشین جزو نشان‌گرهای با اهمیت شناخته شدند که هر دو از نشان‌گرهای بین ژنی بودند. ژن‌های *CNOT10* و *MGAT3* توسط مدل خطی و ژن *LncRNA* توسط دو روش یادگیری ماشین برای صفت وزن بدن در ۹ هفتگی شناسایی شدند. ژن *CNOT10* توسط Emrani و همکاران (۲۰۱۷) برای صفات رشد در جوجه‌های گوشتی گزارش شده است. ژن *MGAT3* (مانوزیل گلیکوزامینیل ترانسفراز ۳) ژنی است که مسئول تولید آنزیم MGAT3 است. این آنزیم در حیوانات و انسان در فرایند گلیکوزیلاسیون، یعنی اضافه کردن گروه‌های گلیکوز به پروتین‌ها، نقش دارد. این فرآیند به پروتین‌ها کمک می‌کند تا ساختار و عملکرد صحیح خود را به دست آورند و در عملکردهای مختلف در سطح سلول و بافت شرکت کنند. این ژن در مطالعات به عنوان ژن تاثیرگذار بر صفات رشد و محدود کننده سرعت جذب چربی در رژیم غذایی در گوساله‌ها گزارش شده است (Lyu *et al*., 2021; Sun *et al* 2012). ژن *LncRNA* در جریان فعالیت ژنتیکی و تنظیم بیان ژن‌ها نقش دارد و چون در کدگذاری پروتین نقش ندارد به-عنوان "غیرکدگذارنده" نامیده می‌شوند. این ژن معمولاً بیشتر از ۲۰۰ نوکلئوتید طول دارد و به صورت تراکم بالا در نواحی ژنوم یافت می‌شود. بطور خاص، برخی از *LncRNA* ها در رشد و تکثیر عضله نقش دارند و می‌توانند در تنظیم بیان ژن‌ها و عملکرد پروتین‌های مرتبط با رشد و تکثیر عضلات مؤثر باشند (Ren *et al*., 2018). نتایج هستی‌شناسی (Gene Ontology) در سطح فرآیندهای بیولوژیکی با استفاده از ۱۰ نشانگر برتر انتخاب شده توسط ۳ روش برای صفات وزن بدن در جدول ۵ گزارش شده است.

جدول ۵. هستی‌شناسی و فرآیند بیولوژی با استفاده از ۱۰ نشانگر برتر تعیین شده با روش های رگرسیون خطی، جنگل تصادفی و گرادیان بوستینگ

نام ژن	p-value	توصیف هستی‌شناسی	شناسه هستی‌شناسی
رگرسیون خطی			
CACNA1I, PPP1R12A, GPR85, RARB, IGF1	0.008	signal transduction	GO:0007165
THRB, RARB	0.012	retinoic acid receptor signaling pathway	GO:0048384
THRB, RARB	0.029	hormone-mediated signaling pathway	GO:0009755
جنگل تصادفی			
ARHGAP32, TAGAP, KALRN	0.005	regulation of small GTPase mediated signal transduction	GO:0051056
MAP2, CAMSAP2, CDK5RAP2	0.007	microtubule cytoskeleton organization	GO:0000226
MAP2, CAMSAP2	0.011	regulation of microtubule polymerization	GO:0031113
BRINP1, MAP2	0.018	central nervous system neuron development	GO:0021954
MAP2, CDK5RAP2	0.030	microtubule bundle formation	GO:0001578
KCNJ5, KCNJ1	0.040	potassium ion import across plasma membrane	GO:1990573
BRINP1, FZD7	0.064	cellular response to retinoic acid	GO:0071300
KCNJ5, KCNJ1	0.069	potassium ion transport	GO:0006813
SGO1, CDK5RAP2	0.085	chromosome segregation	GO:0007059
گرادیان بوستینگ			
RER1, PLPP3	0.038	retrograde vesicle-mediated transport, Golgi to ER	GO:0006890
FZD7, PLPP3	0.075	canonical Wnt signaling pathway	GO:0060070
SPEN, TIMP4	0.087	Notch signaling pathway	GO:0007219

ژن‌های گزارش شده توسط روش جنگل تصادفی فرآیندهای بیولوژیکی بیش‌تری را نسبت به دو روش دیگر نشان دادند. مسیرهای بیولوژیکی ژن‌های کشف شده توسط این روش، نقش‌های بسیار مهمی در سیستم عصبی، استحکام اسکلت سلولی، تقسیم سلولی و کنترل‌کننده سیستم‌های میکروتوبولی دارند که این فرآیندها می‌توانند وزن بدن را تحت تاثیر قرار دهند (Fehm *et al* 2014). از جمله این ژن‌ها می‌توان به دو ژن *MAP2* و *CAMSAP2* اشاره کرد که با ۳ نشانگر برتر شناسایی شده توسط روش‌های یادگیری ماشین مطالعه حاضر مرتبط بوده و در بیشتر فرآیندهای بیولوژیکی بدن شرکت دارند. مشاهده نشان‌گرهای با اهمیت و ژن‌های مرتبط با آن‌ها توسط روش‌های یادگیری ماشین در مقابل با مدل خطی نشان می‌دهد که ممکن است برخی از نشان‌گرها و ژن‌های تاثیرگذار بر صفات به دلیل فرضیات موجود در مدل‌های خطی کشف نشوند. در مطالعات به تاثیر تعداد کم نمونه‌ها بر افزایش نرخ مثبت کاذب نتایج حاصل از مدل‌های خطی اشاره شده است (Hong and Park, 2012). بنابراین استفاده از رویکردهای ترکیبی و مقایسه مدل‌های مختلف می‌تواند راه‌کار مناسبی برای پیدا کردن نشان‌گرهای مرتبط با صفات اقتصادی باشد.

نتیجه گیری و پیشنهادها

در مطالعه حاضر، پویش ژنومی صفات وزن بدن در جوجه‌های گوشتی در سنین ۶ و ۹ هفتگی با روش‌های جنگل تصادفی، گرادیان بوستینگ، و مدل خطی ساده، جهت شناسایی نشان‌گرهای مهم با استفاده از تراشه 60K چندشکلی‌های تک‌نوکلئوتیدی انجام شد. نشان‌گرهای مهمی از جمله GGaluGA141221 و GGaluGA142838 توسط مدل خطی و نشانگرهای GGaluGA308573، Gga_rs13743072، Gga_rs13614212، GGaluGA322130 و Gga_rs15763229 توسط روش‌های یادگیری ماشین برای صفات وزن بدن در ۶ و ۹ هفتگی شناسایی شدند. نشان‌گرهای شناسایی شده در روش‌های یادگیری ماشین با ژن‌های جدیدی از جمله *MAP2*، *ACSL1*، *CAMSAP2*، *FAM117B*، *SLC4A4*، *TIMP4* و *LncRNA* مرتبط بودند. مطالعات هستی‌شناسی ژن نشان داد که ژن‌های کشف شده توسط روش جنگل تصادفی در عملکردهای بیولوژیکی متفاوت و مهمی نقش داشتند که قبلا در جمعیت مورد مطالعه توسط مدل خطی گزارش نشده بودند. بنابراین به نظر می‌رسد به جای استفاده از مدل‌های آماری رایج مانند رگرسیون خطی به تنهایی، استفاده از روش‌های یادگیری ماشین و حتی ادغام روش‌ها با یکدیگر می‌تواند در تجزیه و تحلیل دقیق‌تر پویش ژنومی کمک‌کننده باشد.

سپاسگزاری

بدین‌وسیله، نگارندگان مقاله، از پروفسور Just Jensen از دانشگاه Aarhus دانمارک برای تعیین ژنوتیپ‌ها و تامین هزینه‌های مالی آن، و از شرکت کاب ونترس برای در اختیار قرار دادن تراشه ایلومینا ۶۰k قدردانی می‌کنند.

منابع

Arabnejad, M., Montgomery, C. G., Gaffney, P. M., McKinney, B. A. (2020). Nearest-neighbor projected distance regression for epistasis detection in GWAS with population structure correction. *Frontier Genetics*, 11:784.

Bahadoran, S., Dehghani Samani, A., & Hassanpour, H. (2018). Effect of heat stress on the gene expression of ion transporters/channels in the uterus of laying hens during eggshell formation. *Stress*, 21(1), 51-58.

Baştanlar, Y., & Özuysal, M. (2014). Introduction to machine learning. *Methods in Molecular Biology*, 1107: 105-128.

Boulesteix, A. L., Janitza, S., ruppa, J. K., & König I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Technical Report. *Department of Statistics*, University of Munich.

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

Brew, K., Dinakarpandian, D., & Nagase, H. (2000). Tissue inhibitors of metalloproteinases: evolution, structure and function. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 1477(1-2), 267-283.

Cha, J., Choo, H., Srikanth, K., Lee, S. H., Son, J. W., Park, M. R., Kim, N., Jang, G. W., & Park, J. E. (2021). Genome-wide association study identifies 12 loci associated with body weight at age 8 weeks in Korean native chickens. *Genes*, 12(8), 1170.

Cremonesi, P., Capoferri, R., Pisoni, G., Del Corvo, M., Strozzi, F., Rupp, R., Caillat, H., Modesto, P., Moroni, P., Williams, J. L., & Castiglioni, B. (2012). Response of the goat mammary gland to infection with Staphylococcus aureus revealed by gene expression profiling in milk somatic and white blood cells. *BMC Genomics*, 13(1), 1-17.

Dadousis, C., Somavilla, A., Ilska, J. J., Johnsson, M., Batista, L., Mellanby, R. J., Headon, D., Gottardo, P., Whalen, A., Wilson, D., & Dunn, I. C. (2021). A genome-wide association analysis for body weight at 35 days measured on 137,343 broiler chickens. *Genetics Selection Evolution*, 53, 1-14.

Dzomba, E. F., Chimonyo, M., Snyman, M. A., & Muchadeyi, F. C. (2020). The genomic architecture of South African mutton, pelt, dual- purpose and nondescript sheep breeds relative to global sheep populations. *Animal Genetics*, 51 (6), 910-923.

Emrani, H., Vaez Torshizi, R., Masoudi, A. A., & Ehsani, A. (2017). Identification of new loci for body weight traits in F2 chicken population using genome-wide association study. *Livestock Science*, 206, 125–131.

Enoma, D. O., Bishung, J., Abiodun, T., Ogunlana, O., & Osamor, V. C. (2022). Machine learning approaches to genome-wide association studies. *Journal of King Saud University-Science*, *34*(4), 101847.

Fehm, L., Kern, W., & Peters, A. (2004). Body weight regulation through the central nervous system. The development of a pathogenetically based adiposity therapy. *Medizinische Klinik*, 99 (11), 674-679.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38, 367–378

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29 (5), 1189-1232.

Goddard M. E., & Hayes, B. J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10, 381–391.

Goldstein, B. A., Hubbard, A. E., Cutler A., & Barcellos L. F. (2010). An application of random forests to a genome-wide association dataset: Methodological considerations and new findings. *BMC Genetics*, 11, 49.

Harada, A., Teng, J., Takei, Y., Oguchi, K., & Hirokawa, N. (2002). MAP2 is required for dendrite elongation, PKA anchoring in dendrites, and proper PKA signal transduction. *The Journal of Cell Biology*, 158 (3), 541-549.

Hayes B. (2013). Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). *Methods in Molecular Biology*, 1019, 149-69.

Hong E. P. & Park J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics Inform*, 10(2), 117-22.

Hrabia, A., Miska, K. B., Schreier, L. L., Proszkowiec-Weglarz, M. (2022) Altered gene expression of selected matrix metalloproteinase system proteins in the broiler chicken gastrointestinal tract during post-hatch development and coccidia infection. *Poultry Science*, 101(8):101915.

Hu, T., Darabos, C., & Urbanowicz, R. (2020). Machine learning in genome-wide association studies. *Frontiers in Genetics*, *11*, 593958.

Jin, C. F., Chen, Y. J., Yang, Z. Q., Shi, K., & Chen, C. K. (2015). A genome-wide association study of growth trait-related single nucleotide polymorphisms in Chinese Yancheng chickens. *Genetics and Molecular Research*, 14 (4), 15783-15792.

Kanakachari, M., Ashwini, R., Chatterjee, R. N., & Bhattacharya, T. K. (2021). Transcriptome analysis reveals potential mechanisms and pathways underlying embryonic development with respect to muscle growth and egg production in slow and fast growing chickens. *BMC Genomics*, 13 (1), 58.

Li, B., Zhang, N., Wang, Y. G., George, A. W., Reverter, A., & Li, Y. (2018). Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Frontiers in Genetics*, 9, 237.

Liaw, A., & Wiener, M. (2015). randomForest: Breiman and Cutler's random forests for classification and regression. *R package version*, *4*, 14.

Liu, L., Cui, H., Xing, S., Zhao, G., & Wen, J. (2019). Effect of divergent selection for intramuscular fat content on muscle lipid metabolism in chickens. *Animals*, 10(1), 4.

Lyu, S., Yang, P., Liu, Y., Song, T., Zhang, Z., Shi, Q., Chen, F., Liu, X., Li, Z., Ru, B., & Cai, C. (2021). Genetic effects of MOGAT1 gene SNP in growth traits of Chinese cattle. *Gene*, 769, 145201.

Mebratie, W., Madsen, P., Hawken, R. Rome, H., Marois, D., Henshall, J., Bovenhuis, H., & Jensen J. (2019). Genetic parameters for body weight and different definitions of residual feed intake in broiler chickens. *Genetics Selection Evolution*, 51, 53.
Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*. 11,110.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81 (3), 559-575.
Ren, T., Li, Z., Zhou, Y., Liu, X., Han, R., Wang, Y., Yan, F., Sun, G., Li, H., & Kang, X. (2018). Sequencing and characterization of lncRNAs in the breast muscle of Gushi and Arbor Acres chickens. *Genome*, 61(5), 337-347.
Resnyk, C. W., Carré, W., Wang, X., Porter, T. E., Simon, J., Le Bihan-Duval, E., Duclos, M. J., Aggrey, S. E., & Cogburn, L.A. (2017). Transcriptional analysis of abdominal fat in chickens divergently selected on bodyweight at two ages reveals novel mechanisms controlling adiposity: validating visceral adipose tissue as a dynamic endocrine and metabolic organ. *BMC Genomics*, 18, 1-31.
Ridgeway, G. (2013). Package'GBM': Generalized Boosted Regression Models. *R Package version*, 2.
Srikanth, K., Lee, S. H., Chung, K. Y., Park, J. E., Jang, G. W., Park, M. R., Kim, N. Y., Kim, T. H., Chai, H. H., Sun, J., Zhang, C., Lan, X., Lei, C., & Chen, H. (2012). Exploring polymorphisms and associations of the bovine MOGAT3 gene with growth traits. *Genome*, *55*(1), 56-62.
Sun, S., Dong, B., & Zou, Q. (2021). Revisiting genome-wide association studies from statistical modelling to machine learning. *Briefings in Bioinformatics*, *22*(4), 263.
Teuliere, J., Cordes, S., Singhvi, A., Talavera, K., & Garriga, G. (2014). Asymmetric neuroblast divisions producing apoptotic cells require the cytohesin GRP-1 in *Caenorhabditis Elegans*. *Genetics*, 198(1), 229-247.
Tian, W., Wang, D., Wang, Z., Jiang, K., Li, Z., Tian, Y., Kang, X., Liu, X., & Li, H. (2021). Evolution, expression profile, and regulatory characteristics of ACSL gene family in chicken (Gallus Gallus). *Gene*, 764, 145094.
Wang, J., Yuan, X., Ye, S., Huang, S., He, Y., Zhang, H., Li, J., Zhang, X., & Zhang, Z. (2019). Genome wide association study on feed conversion ratio using imputed sequence data in chickens. *Asian-Australasian Journal of Animal Sciences*, 32 (4), 494-500.
Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., & Beckmann, J.S. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45 (11), 1274-1283.
Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14 (7), 507–515.
Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., & Girón, C. G. (2016). Ensembl 2016. *Nucleic Acids Research*, 44, 710-716.
Yue, Q., Chen, Y., Chen, H., & Zhou, R. (2022). Transcriptome profile reveals novel candidate genes associated with bone strength in end-of-lay hens. *Animal Biotechnology*, 1-9.
Zhang, G. X., Fan, Q. C., Zhang, T., Wang, J. Y., Wang, W. H., Xue, Q., & Wang, Y. J. (2015). Genome-wide association study of growth traits in the Jinghai Yellow chicken. *Genetics and Molecular Research*, 14(4), 15331-15338.

Extended Abstract

Introduction

By employing Genome-Wide Association Studies (GWAS) and identifying Single Nucleotide Polymorphisms (SNPs) and their associations with genes influencing traits, the necessary molecular information for marker and gene selection for improving quantitative traits has been provided. However, in GWAS studies, the most common approach used to examine the effects of markers is simple regression and P-value, which disregards issues such as increasing the false positive rate, overestimation of marker effects, ignoring linkage disequilibrium between markers, incorrect assumption of independence among markers, assumption of all genomic variables following a normal distribution, and neglecting the interaction effects between the SNP markers. Consequently, a suitable alternative to address these problems is conducting GWAS based on machine learning methods. The main objective of this study is to identify important and influential markers for body weight traits measured at 6 and 9 weeks of age in F_2 chickens population using the RF and GB methods and comparing them with the linear model (LM) approach.

Material and methods

For the current study, body weight data at ages 6 and 9 weeks of 312 F_2 chickens, resulting from two-way crossbreeding between fast-growing commercial Arin line (R) and indigenous fowls from West Azerbaijan province (O) were used. At age of 70 days, DNA from blood samples of chickens was extracted and stored at -20°C. These DNAs were used to identify genotype of each bird, using the 60k Illumina Chicken SNP BeadChip, containing 54,340 SNP markers provided by Cobb Vantress with cooperation of Arhus University of Denmark. The phenotypic data were adjusted for sex and hatching effects, and three methods including linear model, random forest, and gradient boosting were used to identify the important markers. The top ten markers for body weight traits were identified for each method. In addition, genes located within the Mb-1 region above and below the three top markers, as identified by each method in the genomic region, were determined using the NCBI and Ensemble databases from the reference genome of chicken (Gallus Gallus).

Results and discussion

Identification of important markers and the corresponding genes using machine learning methods showed that some markers and genes influencing traits might not be identified by linear model, indicating that machine learning methods including random forest and gradient boosting were suitable tools for selecting important markers. These associated markers were GGaluGA308573, GGaluGA255033, Gga_rs13614212, Gga_rs13743072, GGaluGA258772, Gga_rs14034395, and Gga_rs13858398. By examining only 3 of the most important markers in each method, new related genes such as *MAP2*, *ACSL1*, *CAMSAP2*, *FAM117B*, *SLC4A4*, *TIMP4*, and *LncRNA* were identified which were not detected by linear model in previous studies. Literature results reported that these genes are regulating microtubule-stabilizing activity, cell's shape, intestinal tissues during post-hatch development, augmenting adipogenesis, tissue growth and morphogenesis, fatty acid metabolism and abdominal fat deposition, intramuscular fat content, meat tenderness, and flavor, axon development, dendrite development, organelle organization, bicarbonate secretion and absorption and intracellular pH, and induction of apoptosis (Programmed death of cells) in chickens.

Conclusion

In this study, random forest, gradient boosting, and simple linear models were used to detect important markers association with body weight traits at 6 and 9 weeks of ages in broiler chickens using a 60k Chicken SNP BeadChip. The most important markers that were identified by linear model were GGaluGA141221, GGaluGA142838. For machine learning methods, the top markers were GGaluGA308573, Gga_rs13743072, Gga_rs13614212, GGaluGA322130, and Gga_rs15763229. These markers were associated with genes that control several biochemical, physiological and biological functions in chickens. Results indicate that machine learning algorithms were able to identify new genes for body weight traits that were not previously identified by linear model.