

مقاله علمی-پژوهشی

مدل‌سازی و پیش‌بینی عملکرد گندم براساس متغیرهای هواشناسی با استفاده از روش‌های ترکیبی هوش مصنوعی

امین امیرعشایری^۱، وحید رضاوردی‌نژاد^۲، جواد بهمنش^{۳*}، فرخ اسدزاده^۴، مینا رحیمی^۵

تاریخ دریافت: ۱۴۰۳/۱۰/۲۵ تاریخ پذیرش: ۱۴۰۴/۰۱/۱۲

چکیده

گندم یک نقش اساسی در امنیت غذایی کشور داشته و تخمین میزان عملکرد آن، کمک موثری در تصمیم‌گیری‌های کلان و منطقه‌ای دارد. برای پیش‌بینی عملکرد محصول، روش‌های هوش مصنوعی به عنوان یکی از ابزارهای بسیار مناسب شناخته شده‌اند. لذا در این تحقیق طیف وسیعی از مدل‌های یادگیری ماشین که شامل شبکه عصبی مصنوعی (ANN)، جنگل تصادفی (RF)، رگرسیون بردار پشتیبان (SVR) و اسپیلاین رگرسیون تطبیقی چند متغیره (MARS) در استان آذربایجان غربی که یکی از مناطق بسیار مهم در تولید گندم کشور است، مورد ارزیابی قرار گرفتند. براساس متغیرهای مؤثر بر عملکرد گندم شامل مجموع بارش فصل زراعی، شمار روز اولین و آخرین بارش مؤثر ۱۰ میلی‌متری، دما، رطوبت نسبی، تبخیر، میانگین ساعات آفتابی و به همراه مجموع تعداد روزهای یخبندان به عنوان متغیرهای ورودی در نظر گرفته شدند. برای بهینه کردن پارامترهای ورودی مؤثر، از روش آزمون گاما در ترکیب با الگوریتم ژنتیک استفاده شد. با استفاده از آزمون گاما و الگوریتم ژنتیک ۸ متغیر برای ورودی مدل‌ها انتخاب گردید. همچنین برای بهبود دقت پیش‌بینی مدل‌ها از ترکیب آن‌ها با الگوریتم گروهی تجزیه مد تجربی (EEMD) استفاده شد. نتایج تخمین‌ها نشان داد که مدل EEMD-MARS نتایج مناسب‌تری ارائه داد، که معیارهای ارزیابی خطا شامل: $RMSE=0.112(\text{ton}\cdot\text{ha}^{-1})$ ، $NSE=0.945$ ، $SI=0.101$ و $MAE=0.088(\text{ton}\cdot\text{ha}^{-1})$ برای مرحله آزمون بدست آمد. همچنین ۱۴ تابع مختلف برای تخمین عملکرد گندم در این مدل استخراج گردید. با این حال عملکرد مدل EEMD-SVR با نماهای ارزیابی خطا شامل: $RMSE=0.132(\text{ton}\cdot\text{ha}^{-1})$ ، $MAE=0.080(\text{ton}\cdot\text{ha}^{-1})$ و $NSE=0.923$ برای مرحله آزمون عملکرد مناسب‌تری داشت.

واژه‌های کلیدی: آزمون گاما، الگوریتم گروهی تجزیه مد تجربی، محصولات دیمی، یادگیری ماشین

مقدمه

راه‌کارهایی برای غلبه بر کمبود آب در این مناطق همواره از اولویت‌های مطالعاتی و تحقیقاتی است (سیفی و همکاران، ۱۳۸۹). متأسفانه عملکرد فعلی مزارع گندم بسیار پایین‌تر از میانگین جهانی است. در ایران گندم با سطح زیر کشت بیش از ۶ میلیون هکتار حدود نیمی از مساحت زمین‌های زراعی (۶/۵۶ درصد) را به خود اختصاص داده است که ۳۴/۲ درصد به صورت کشت آبی و ۶۵/۸ درصد این اراضی به صورت کشت دیم است (جعفری و همکاران، ۱۴۰۲). کشت دیم که نقش مهمی را در تأمین غذای مردم ایفا می‌کند، دارای سابقه‌ی طولانی است که میزان تولید محصولات به‌ویژه گندم در دیم‌زار صرف نظر از ساختمان نوع محصول به مدیریت صحیح و بهره‌وری بهینه عوامل مختلف محیطی به‌ویژه شرایط آب و هوایی بستگی دارد. در این میان تغییرات مکانی و نوسان‌های سال به سال و یا دوره‌ای شرایط اقلیمی، نقش بسیار مهمی در میزان برداشت گندم دیم دارند. بنابراین شناخت و آگاهی از شرایط اقلیمی هر منطقه می‌تواند کمک

ایران کشوری در زمره مناطق خشک و نیمه خشک جهان با متوسط ریزش‌های جوی ۲۵۰ میلی‌متر طبقه‌بندی می‌گردد (Taheri et al., 2019). از عمده‌ترین مشکلات بخش کشاورزی در کشور محدودیت منابع آب و نبود آب با کیفیت مناسب است؛ بنابراین یافتن

۱- دانشجوی دکتری آبیاری و زهکشی، گروه مهندسی آب، دانشگاه ارومیه، ارومیه، ایران

۲- استاد گروه مهندسی آب، گروه مهندسی آب، دانشگاه ارومیه، ارومیه، ایران

۳- استاد گروه مهندسی آب، گروه مهندسی آب، دانشگاه ارومیه، ارومیه، ایران

۴- استاد گروه علوم خاک، گروه علوم خاک، دانشگاه ارومیه، ارومیه، ایران

۵- دانش‌آموخته دکتری گروه مهندسی آب، گروه مهندسی آب، دانشگاه ارومیه، ارومیه، ایران

*- نویسنده مسئول: (Email: j.behmanesh@urmia.ac.ir)

ایالات آمریکا در بازه‌ی زمانی ۲۰۱۶ تا ۲۰۱۸ استفاده کردند که نتایج حاکی از این بود که روش RNN-CNN به‌طور قابل ملاحظه‌ای از سایر روش‌های به‌کار گرفته بهتر عمل کرد (Khaki et al., 2020). کومار و همکاران از تکنیک‌های SVR^{۱۱}، MT^{۱۲}، RF^{۱۳} و GB^{۱۴} و پارامترهای میانگین بارندگی، میزان سموم دفع آفات و میانگین دمای سالانه برای پیش‌بینی عملکرد تولید گندم دیم در مناطقی از هند استفاده کردند. نتایج حاکی از دقت بیشتر مدل جنگل تصادفی با حداقل میانگین خطای مطلق نسبت به روش‌های دیگر بود (Kumar et al., 2022). گرینبرگ و همکاران، موراکامی و همکاران و معراج و همکاران برای پیش‌بینی عملکرد گندم از روش‌های یادگیری ماشین در مناطق مختلف جهان استفاده نمودند (Grinberg et al., 2020; Murakami et al., 2021; Meraj et al., 2022). به عنوان نمونه، آفرین و همکاران با استفاده از مدل ماشین بردار پشتیبان عملکرد محصول برنج را برای مناطق محصولات کشاورزی بنگلادش پیش‌بینی نمودند (Afrin et al., 2018). در تحقیق دیگری، گاندهی و همکاران، پیش‌بینی عملکرد برنج و عوامل موثر بر عملکرد آن را با استفاده از شبکه‌ی عصبی مصنوعی^{۱۴} (MLP) با پارامترهای ورودی بارش، حداقل دما، میانگین دما، حداکثر دما و تبخیر-تعرق مرجع محصول برای سال‌های ۱۹۹۸ تا ۲۰۰۲ بررسی کردند. مدل MLP با داشتن میانگین خطای کم‌تر به عنوان مدل برتر برای پیش‌بینی مناطق مختلف ایالت ماهاراشترا در هند انتخاب شد (Gandhi et al., 2016). سونیل و همکاران، از روش‌های طبقه‌بندی ANN^{۱۵} و K-NN^{۱۶} برای عملکرد محصولات برنج، سیب‌زمینی و گندم استفاده نمودند. ایشان برای این منظور داده‌های عملکرد محصول را در مناطق بنگلادش در نظر گرفتند. نتایج نشان داد که روش ANN پیش‌بینی بهتری برای سیب‌زمینی و برنج و روش LR پیش‌بینی بهتری برای گندم دارند (Sunil et al., 2022). در تحقیقی دیگر عباس و همکاران از چهار الگوریتم (ML) شامل^{۱۷} (EN)، (KNN)، (SVR) و (LR) برای پیش‌بینی عملکرد محصول در مقیاس مزرعه استفاده نمودند (Abbas et al., 2020). استفاده از روش‌های ترکیبی پیش‌پردازنده با روش‌های یادگیری ماشین باعث افزایش دقت مدل‌ها می‌شود که محققین زیادی از این تکنیک‌های پیش‌پردازش‌کننده برای استخراج ویژگی‌های قرار داده شده در داده‌های سری زمانی مانند تبخیر-تعرق مرجع روزانه (امیرعشایری و همکاران، ۱۳۹۹)،

بسیار زیادی به کشاورزان در راستای افزایش بهره‌وری محصول داشته باشد (جلالی و همکاران، ۱۳۹۶). استفاده بهینه و مناسب از منابع آبی و کشاورزی برای افزایش بازده محصولات نیازمند درک دقیق از پارامترهای مهم هواشناسی و آب و هوایی و بخصوص بارش در مناطق دیم، و سیستم‌های هیدرولوژیکی هر حوضه در منطقه می‌باشد. بارش یکی از مهم‌ترین پارامترها در هیدرولوژی و سیستم‌های هواشناسی است (Dastorani et al., 2016). پیش‌بینی عملکرد محصولات کشاورزی امروزه یک عامل مهم و کلیدی در پیشرفت حوزه‌ی کشاورزی است، همچنین امروزه به دلیل در دسترس بودن داده‌های بسیار زیاد در کشاورزی، کارهای تحقیقاتی زیادی با استفاده از تکنیک‌های مختلف برای پیش‌بینی، ارزیابی و مدل‌سازی محصولات در حال انجام است (جعفری و همکاران، ۱۴۰۲). روش‌های یادگیری ماشین^۱ (ML) و داده‌کاوی بیش‌ترین کاربرد را برای پیش‌بینی تولید محصولات کشاورزی و فاکتورهای تأثیرگذار بر تولید بهینه محصولات در مسائل کشاورزی را دارند. زیرا الگوریتم‌های یادگیری ماشین بهتر از روش‌های آماری کلاسیک برای کشف روابط غیرخطی پنهان عمل می‌کنند (Arshad et al., 2023). این روش‌ها برای استخراج اطلاعات مهم و مفید از یک مجموعه داده بزرگ استفاده می‌شوند که برای پیش‌بینی عملکرد محصولات از این روش‌ها می‌توان با استخراج ویژگی اطلاعات از یک مجموعه داده و تبدیل به یک ساختار قابل فهم استفاده کرد. در سال‌های اخیر، استفاده از روش‌های هوش مصنوعی نظیر شبکه عصبی مصنوعی^۲ (ANN)، ماشین‌های بردار پشتیبان^۳ (SVM) و برنامه‌ریزی بیان ژن^۴ (GEP) برای مدل‌سازی و حل مسائل در مهندسی آب (داده‌های هیدرولوژیکی کافی را دارا باشند) به علت مزایای قابل تأمل آن، فراگیر شده است (Rezaie-Balf et al., 2019). همچنین روش‌های یادگیری مبتنی بر هوش مصنوعی مانند جنگل تصادفی^۵ (RF)، شبکه‌های عصبی مصنوعی، شبکه‌های عصبی عمیق^۶ و ماشین‌های بردار پشتیبان (SVM) به دلیل توانایی آنها در مقابله با مسائل پیچیده و غیرخطی در مطالعات مختلف کشاورزی استفاده شده‌اند (Kim et al., 2019; Sinwar et al., 2020).

خاکی و همکاران برای بررسی مطالعات مختلف کشاورزی از روش‌های CNN^۷، RNN^۸، RF^۹، DFF^{۱۰} و LASSO^{۱۱} در ۱۳ ایالات

- 1- Machine Learning
- 2- Artificial Neural Network
- 3- Support Vector Machine
- 4- Gene Expression Programing
- 5- Random Forest
- 6- Deep Neural Networks
- 7- Convolution Neural Networks
- 8- Recurrent Neural Networks
- 9- Deep Feed Forward Neural Networks
- 10- Least Absolute Shrinkage and Selection Operator

11- Support Vector Regression

12- Model Tree

13- Gradient boosting

14- Multi-Layer Perceptron

15- Linear Regression

16- k-Nearest Neighbors algorithm

17- Elastic Net

دارد. این استان از نظر موقعیت جغرافیایی از شمال به جمهوری آذربایجان و ترکیه، از غرب به کشورهای ترکیه و عراق، از شرق به استان آذربایجان شرقی و استان زنجان و از جنوب به استان کردستان محدود است. مساحت استان بدون احتساب دریاچه ارومیه برابر ۳۷۰۵۹ کیلومتر مربع است که سیزدهمین استان بزرگ کشور محسوب می‌شود، و ۲/۲۵ درصد مساحت کل کشور با ۱۹ شهرستان را تشکیل می‌دهد. در شکل (۱) موقعیت جغرافیایی استان آذربایجان غربی و پراکنش ایستگاه‌های سینوپتیک مورد استفاده در این پژوهش نشان داده شده است. این استان با ارتفاع متوسط ۱۳۰۰ متر از سطح دریاهای آزاد یکی از مناطق کوهستانی کشور بوده و بر اساس ساختار طبیعی استان، اکوسیستم‌های ویژه‌ای از ترکیب گیاهان در سطوح مختلف پوشش گیاهی به وجود آمده است. دریاچه ارومیه که در شرق استان واقع است، در شمال غرب ایران بیستمین دریاچه جهان از لحاظ وسعت محسوب می‌گردد. وسعت حوضه آبریز دریاچه ۵۲۳۵۵ کیلومتر مربع است که از این مقدار حدود ۵۸۲۲ کیلومتر مربع وسعت خود دریاچه مستقیماً با ارتفاع آب دریاچه ارتباط دارد. تراز کف دریاچه از سطح آزاد دریاها ۱۲۶۸ متر است. متوسط درجه حرارت سالیانه آن ۱۱ درجه سانتی‌گراد و میانگین سالانه بارندگی در این منطقه بین ۱۸۰ تا ۵۰۰ میلی‌متر متغیر است (گودرزی و همکاران، ۱۳۹۴). در این تحقیق، تمامی شهرستان‌های استان به غیر از شهرستان‌های شوط، پلدشت و شاهین‌دژ به دلیل عدم وجود آمار کافی، مورد بررسی قرار گرفتند.

با توجه به اهداف تحقیق، ابتدا داده‌های هواشناسی مربوط به هر ایستگاه سینوپتیک واقع در هر شهرستان، از استان آذربایجان غربی با طول دوره‌ی آماری ۲۲ سال (سال زراعی ۱۳۸۱-۱۳۸۰ تا ۱۴۰۲-۱۴۰۱) منطبق با دوره‌ی کاشت تا برداشت محصول گندم دیم پاییزه (از اول مهر تا آخر تیر) از سازمان هواشناسی دریافت شد. در هر یک از متغیرهای مذکور نواقص آماری موجود در داده‌های مربوط، با استفاده از روش‌های همبستگی نظیر میانگین‌گیری بازسازی گردید. جدول (۱) مشخصات آماری پارامترهای مورد استفاده شامل، مجموع بارش فصل زراعی (R)، شمار روز اولین بارش مؤثر ۱۰ میلی‌متری (R_{10})، میانگین حداکثر دمای فصل زراعی (T_{max})، میانگین حداقل دمای فصل زراعی (T_{min})، میانگین متوسط دمای فصل زراعی (T_{avg})، میانگین حداکثر رطوبت نسبی فصل زراعی (RH_{max})، میانگین متوسط رطوبت نسبی فصل زراعی (RH_{avg})، و تعداد روزهای یخبندان (SN) را نشان می‌دهد. بیشترین دمای متوسط در محدوده مطالعاتی در فصل زراعی ۱۳/۴ درجه سانتی‌گراد بوده است. تعداد روزهای یخبندان در محدود مطالعاتی در فصل زراعی بین ۲۶ و ۱۶۰ روز بوده و مقدار میانگین آن نیز ۹۶ روز بوده است. میانگین متوسط رطوبت نسبی در محدوده مطالعاتی در فصل زراعی حدود ۵۸ درصد

تبخیر (Deo et al., 2019)، شاخص تشعشع خورشید (Ghaemi et al., 2019)، سرعت باد (Huang et al., 2019)، پارامترهای کیفیت آب (Fijani et al., 2019) و ارتفاع موج دریا (Ali and Prasad., 2019) استفاده کردند. این رویکردها به عنوان تکنیک‌های پیش-پردازنده شامل PCA^1 ، CWT^2 ، $MESA^3$ و MA^4 و الگوریتم‌های های تجزیه مد تجربی EMD^5 می‌باشند. روش تجزیه مد تجربی به عنوان یک روش تجزیه داده‌ها مبتنی بر نویز، توسط هانگ و همکاران (۱۹۹۸) ارائه شد. که سری ورودی فرکانس بالاتر را به مؤلفه‌های (سیگنال) حل شده با فرکانس پایین تر تفکیک می‌کند. در نتیجه با توجه به موارد بیان شده با وجود روش‌های مختلف یادگیری ماشین، اما در مدل‌سازی عملکرد گندم دیم مورد توجه قرار نرفته است و نیازمند بررسی بیشتر می‌باشد. به همین منظور هدف از انجام این تحقیق، ارائه مدل پیش‌بینی، جهت تخمین عملکرد گندم دیمی در استان آذربایجان غربی با استفاده از مدل‌های مختلف یادگیری ماشین می‌باشد. با توجه به اهمیت راهبردی محصول گندم دیم به عنوان محصولی استراتژیک، پیش‌بینی دقیق عملکرد این محصول می‌تواند به بهبود و افزایش امنیت غذایی بیشتر در کشور کمک کند. به دلیل داشتن آب و هوای مناسب برای کشت دیم در این استان هر جا که جنس خاک و شیب زمین مساعد بوده و امکان بهره‌برداری از زمین با آبیاری وجود نداشته باشد، کشت دیم حتی در قطعات کوچک انجام شده است (فرج زاده و زرین، ۱۳۸۱). مدل‌های $MARS^6$ ، ANN ، SVR و RF از جمله مدل‌های در نظر گرفته در این تحقیق بودند. در این تحقیق، الگوریتم پیش‌پردازش‌کننده گروهی $EEMD^7$ با روش مذکور ترکیب شده تا بتوان از طریق این روش ترکیبی، مدلی دقیق تر جهت تخمین مناسب عملکرد گندم داشت. همچنین از دیگر نوآوری‌های این تحقیق، استفاده از ترکیب آزمون گاما با الگوریتم ژنتیک برای شناسایی بهترین پارامترهای ورودی به مدل‌های یادگیری ماشین مورد استفاده می‌باشد.

مواد و روش‌ها

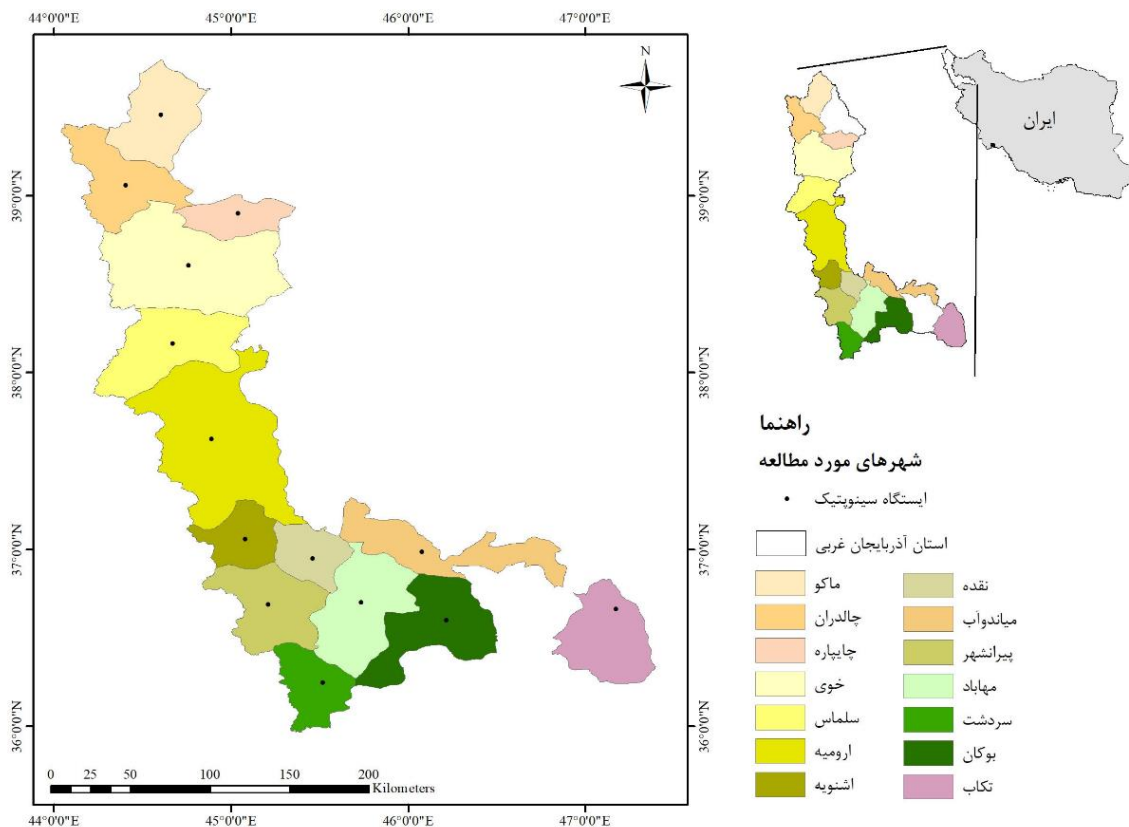
منطقه‌ی مورد مطالعه

منطقه مورد مطالعاتی در این پژوهش استان آذربایجان غربی یکی از استان‌هایی است که در زمینه تولید و مصرف گندم سابقه دیرینه دارد. در کناره شرقی این استان، دریاچه ارومیه قرار

- 1- Principal Component Analysis
- 2- Continuous Wavelet Transform
- 3- Maximum Entropy Spectral Analysis
- 4- Moving Average
- 5- Empirical Mode Decomposition
- 6- Multivariate Adaptive Regression Spline
- 7- Ensemble Empirical Mode Decomposition

ویژگی، ۱- بر عامل خروجی تأثیر داشته باشد، ۲- نسبت به زمان متغیر باشند، هستند. به همین منظور در تحقیق پیش‌رو جهت پیش-بینی عملکرد گندم از پارامترهای ارائه شده در جدول (۱) استفاده شده است.

بوده است. در فصل زراعی بیشترین و کمترین مقدار بارش به ترتیب ۱۳۱۵/۴ و ۱۰۸ میلی‌متر بوده است. همچنین عملکرد گندم دیم بین ۳۱۱۰/۵ و ۲۷۱/۶ کیلوگرم بر هکتار متغیر بوده است. مهم‌ترین نکته در استفاده از مدل‌های یادگیری ماشین انتخاب پارامترهای ورودی مناسب بر عامل خروجی است. انتخاب این پارامترها همواره دارای دو



شکل ۱- موقعیت محدوده‌ی مطالعاتی و ایستگاه‌های سینوپتیک مورد مطالعه

جدول ۱- مشخصات پارامترهای آماری داده‌های مورد استفاده

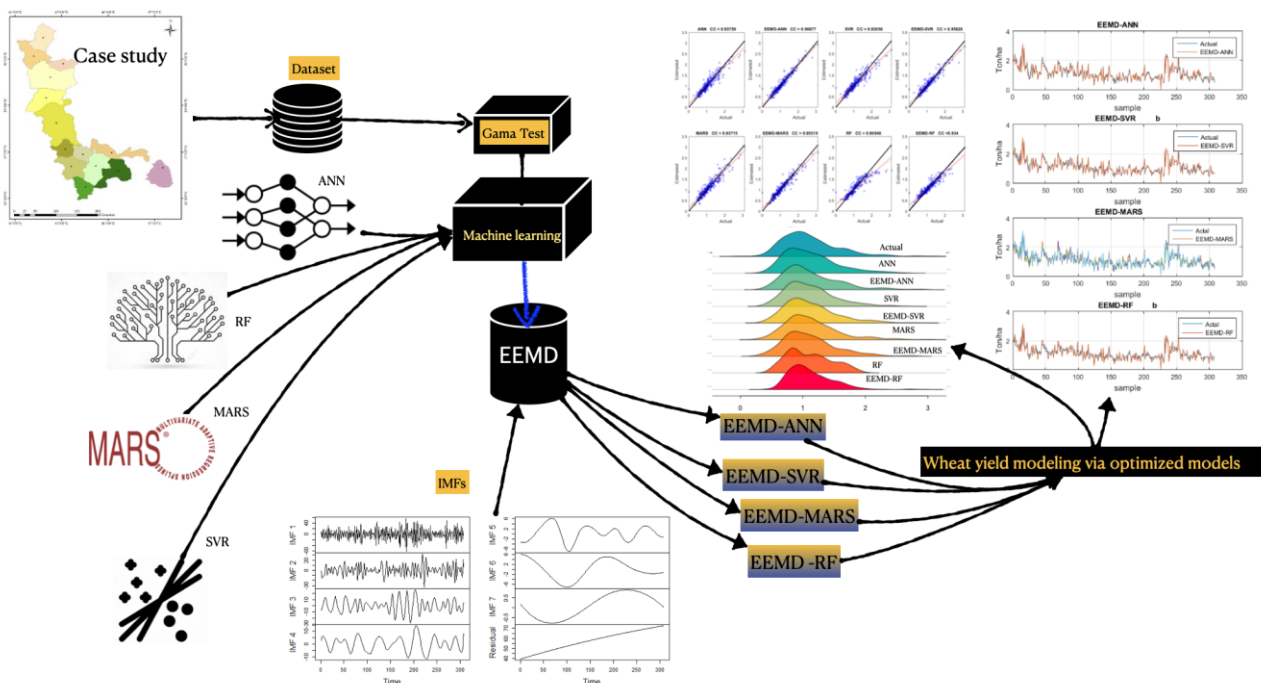
پارامترهای هواشناسی	کمترین	میانگین	بیشترین	انحراف معیار
میانگین حداکثر دمای فصل زراعی (C°)	۱۰/۰	۱۶/۱۲	۲۱/۲۰	۲/۰۱
میانگین حداقل دمای فصل زراعی (C°)	-۰/۲۰	۴/۳۸	۹/۴۰	۲/۰۳
میانگین متوسط دمای فصل زراعی (C°)	۵/۳	۱۰/۱۰	۱۳/۴۰	۱/۷۴
میانگین حداکثر رطوبت نسبی فصل زراعی (%)	۶۰	۷۷	۹۰	۶
میانگین متوسط رطوبت نسبی فصل زراعی (%)	۴۵	۵۸	۶۷	۴
تعداد روزهای یخبندان (day)	۲۶	۹۶	۱۶۰	۲۱
مجموع بارش (mm)	۱۰۸	۳۸۹/۷	۱۳۱۵/۴	۲۰۶/۹
شمار روز اولین بارش مؤثر ۱۰ میلی‌متری (day)	۱	۵۸	۱۴۸	۲۹
عملکرد گندم دیم (kg.ha ⁻¹)	۲۷۱/۶۰	۱۱۱۲/۹۶	۳۱۱۰/۵	۴۴۷/۶۸

شهرستان پرداخته شد. اطلاعات هواشناسی و به همراه عملکرد گندم دیم جمع‌آوری گردید، و ارزیابی روی اطلاعات انجام شد. در ادامه

فلوچارت تحقیق مطابق شکل (۲) در ابتدا به بررسی اطلاعات موجود در هر

در تحقیقات گذشته برای پیش بینی داشتند و به ندرت در این زمینه مورد استفاده قرار گرفته اند پیشنهاد شدند. برای این منظور داده های ورودی به دو بخش آموزش و آزمون به نسبت ۷۵ به ۲۵ و به صورت تصادفی تقسیم شدند. مدل های مذکور با استفاده از داده های آموزش ساخته شده و مقادیر پارامترهای موجود در هر مدل بهینه یابی شدند و سپس عملکرد مدل آموزش دیده شده توسط داده های آزمون مورد ارزیابی مجدد قرار گرفت. در ادامه جهت بهبود نتایج تخمین مدل ها از الگوریتم گروهی تجزیه مد تجزیه استفاده شد. این الگوریتم با هر یک از مدل ها ترکیب شده و موجب کارایی بهتر تخمین عملکرد گندم دیم گردید. در نهایت نتایج حاصل توسط معیارهای ارزیابی خطا و نمودارهای تحلیلی مورد ارزیابی قرار گرفتند. جزئیات بیشتر از روند انجام تحقیق در شکل (۲) قابل مشاهده است.

نواقص موجود در داده ها با استفاده از روش همبستگی بازسازی گردید. در ابتدا با توجه به اطلاعات، ۱۴ متغیر ورودی از آزمون گاما و ترکیب با الگوریتم ژنتیک برای بهینه یابی تعداد ورودی ها استفاده گردید که در نهایت هشت متغیر ورودی شامل، مجموع بارش فصل زراعی، شمار روز اولین بارش مؤثر ۱۰ میلی متری، میانگین حداکثر دمای فصل زراعی، میانگین حداقل دمای فصل زراعی، میانگین متوسط دمای فصل زراعی، میانگین حداکثر رطوبت نسبی فصل زراعی، میانگین متوسط رطوبت نسبی فصل زراعی و تعداد روزهای یخبندان فصل زراعی انتخاب شدند. متغیرهای ورودی انتخاب شده توسط چهار مدل یادگیری ماشین که شامل مدل های MARS، SVR، ANN و RF بود، برای تخمین میزان عملکرد گندم دیم، مورد استفاده قرار گرفتند. این مدل ها از آنجایی که عملکرد مناسبی



شکل ۲- روند انجام مدل سازی عملکرد گندم دیم توسط مدل های یادگیری ماشین

ورودی با ایجاد یک مدل پایدار و هموار نمی توانند محاسبه شوند به- کار رود. در تکنیک GT هر ترکیب ورودی شامل ردیفی از اعداد صفر و یک که ماسک^۲ گفته می شود، می باشد. که عدد صفر نمایانگر عدم مشارکت پارامتر ورودی به مدل و عدد یک نشان دهنده مشارکت آن پارامتر در ترکیب ورودی است (Remesan and Mathew., 2015). جزئیات روش آزمون گاما در مراجعی همچون (Durrant, 2001) ارائه شده است. در این روش انتخاب بهترین ماسک به عنوان

آزمون گاما تست

آزمون گاما اولین بار توسط (Koncar., 1997) ارائه شد. آزمون گاما^۱ (GT) یک مدل غیرخطی بهینه سازی است که با استفاده از آن می توان، ترکیبی مناسب از ورودی های مختلف را برای مدل سازی و ایجاد یک مدل پایدار با کمترین خطا ارائه و بررسی کرد (سالت پور و همکاران، ۱۳۹۲). همچنین این آزمون می تواند تا اندازه های برای تخمین آن بخش از واریانس داده های خروجی که از روی داده های

2- Mask

1- Gamma Test

جنگل تصادفی

جنگل تصادفی یکی از روش‌های یادگیری ماشینی است که برای طبقه‌بندی و رگرسیون استفاده می‌شود. RF گروهی از روش‌ها است، که چندین الگوریتم درخت تصمیم را برای ایجاد پیش‌بینی تکرارشونده از پدیده‌ها ترکیب می‌کند (Breiman, 2001). به طور کلی، عملکرد روش‌های مجموعه‌ای مانند RF، بهتر از مدل‌های رگرسیون پایه فردی است. هدف جنگل تصادفی کاهش واریانس پیش‌بینی است و برای مدل‌سازی اثر غیرخطی متغیرهای ورودی مناسب است (Oshiro et al., 2012). این روش، حساسیتی نسبت به توزیع داده‌ها و وجود نقاط پرت نشان نمی‌دهد. RF می‌تواند ساختارها و مدل‌های پیچیده را به خوبی روابط غیرخطی بین متغیرهای مستقل و وابسته را یاد بگیرد. این روش یکی از الگوریتم‌های یادگیری نظارت شده است که گروهی از درخت تصمیم را به صورت تصادفی ایجاد می‌کند، که توسط روش کیسه‌گذاری^۲ انجام می‌گردد. در ادامه درخت‌های تصمیم با یکدیگر ادغام شده و یک مدل جنگل واحد را برای پیش‌بینی بهتر ایجاد می‌کند. الگوریتم موجود در جنگل تصادفی همواره به دنبال مهم‌ترین ویژگی در هر تقسیم کردن گره‌ها است، که در نتیجه عملکرد مناسبی در تخمین و پیش‌بینی به همراه دارد. برای این تحقیق حدود ۱۰۰ درخت ایجاد شد، که شکل (۳) نمونه‌ای از جنگل تصادفی ساده ایجاد شده برای تحقیق حاضر را نشان می‌دهد. در آن یک درخت به همراه شش گره ایجاد شده است. همان‌طور که مشخص است مقادیر بارش به دو دسته کم‌تر و بیش‌تر از ۳۵۸ تقسیم شده است که موجب توسعه مدل جنگل تصادفی گردیده است. برای جزئیات بیش‌تر در مورد این مدل می‌توانید به تحقیقات (Rigatti, 2017) مراجعه شود.

مدل اسپلاین رگرسیون تطبیقی چند متغیره

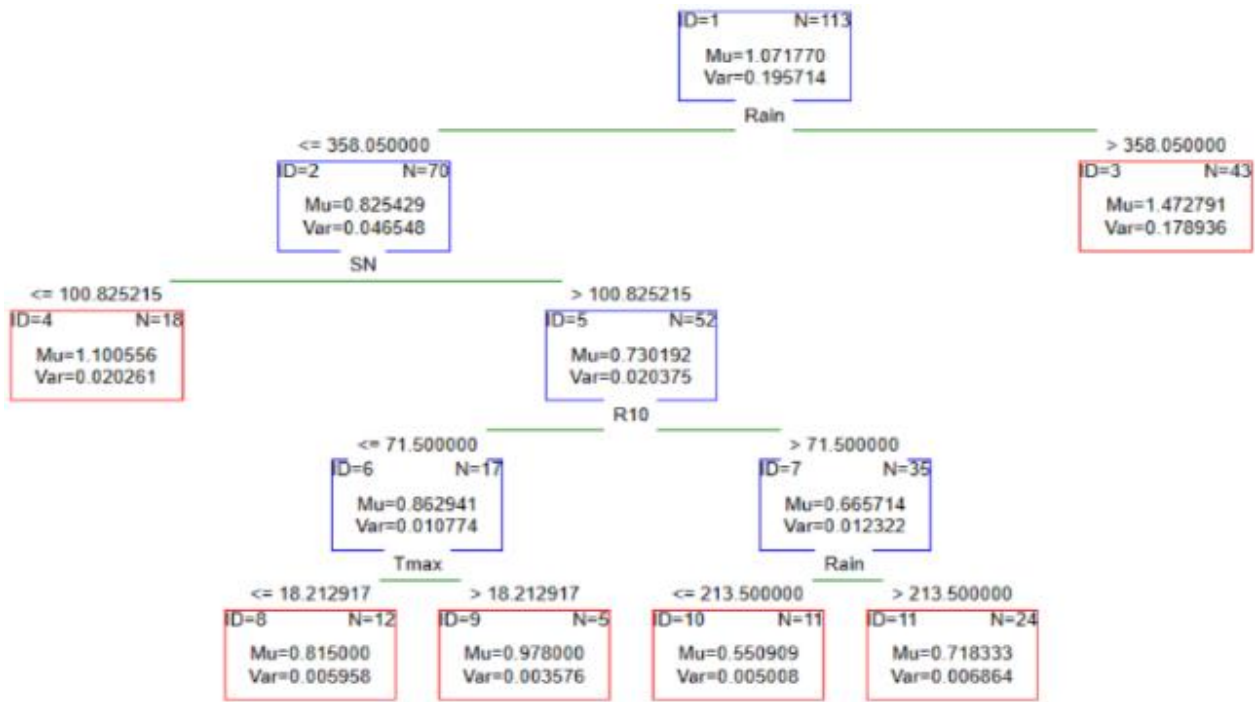
این روش در سال ۱۹۹۱ توسط جرمی فریدمن ارائه شد و یک روش رگرسیونی ناپارامتری است، که برای مدل‌سازی رگرسیونی انعطاف‌پذیر داده‌های با ابعاد بالا ارائه شده است (Friedman, 1991). این روش قادر است روابط غیرخطی پنهان در مجموعه داده‌ها را کشف نماید و بسیار سریع به جواب بهینه برسد (Jamnani et al., 2024). در روش اسپلاین رگرسیون تطبیقی چند متغیره با داشتن متغیر وابسته و مجموعه‌ای از متغیرهای توضیحی مستقل، با استفاده از تبدیلات متغیرهای توضیحی، رابطه غیرخطی متغیر وابسته با متغیرهای توضیحی آشکار می‌شود، بین متغیرهای توضیحی اثرات متقابل تعیین می‌شود و امکان تشخیص متغیرهای مؤثر از متغیرهای غیر مؤثر در مقدار متغیر وابسته به وجود می‌آید. بنابراین مدل کلی MARS را می‌توان به صورت رابطه (۱) بیان کرد:

بهترین پارامترهای ورودی در آزمون گاما بر اساس محاسبه‌ی پارامترهای آماره گاما (Γ) و شاخص (V_{ratio}) انجام می‌شود. جهت تعیین بهترین ماسک نیاز است تمام ترکیبات معنادار (ماسک‌ها) از داده‌های ورودی به صورت دستی برای محاسبه‌ی آماره گاما (Γ) و شاخص (V_{ratio}) ایجاد شود (میرعربی و همکاران، ۱۳۹۷). هرچه آماره گاما و شاخص (V_{ratio}) کمتر و به صفر نزدیک شود بیانگر دقت آن ماسک جهت استفاده در مدل‌سازی می‌باشد و هرگاه به یک نزدیک شوند نشان‌گر خطای تصادفی بسیار بزرگ در مدل‌سازی است و ایجاد مدل بر اساس ماسک‌های انتخاب شده باعث کاهش دقت مدل می‌باشد (قبائی سوق و همکاران، ۱۳۸۹). برای جزئیات بیش‌تر در مورد این آزمون می‌توانید به تحقیقات (Kemp et al., 2004) مراجعه کنید.

مدل‌های یادگیری مورد استفاده

شبکه عصبی مصنوعی

شبکه عصبی مصنوعی در سال ۱۹۹۸ به منظور بهبود نتایج پیش‌بینی‌ها و تخمین توابع معرفی گردید. که جایگزین مناسبی برای معادلات رگرسیونی و معادلات تجربی بود (Azar et al., 2023). این مدل‌ها یک سیستم پردازش اطلاعات توزیع شده موازی هستند که شبیه به شبکه‌های عصبی انسانی عمل می‌کنند (Yao, 1999). بنابراین همانند نورون‌های طبیعی در شبکه عصبی نیز نورون‌ها وظیفه انتقال اطلاعات را دارند. روند کلی این روش چنین است که پس از انتخاب ورودی‌های مدل، ابتدا باید تعداد لایه‌های مخفی و خروجی به همراه تعداد نورون‌ها در هر لایه مشخص شود. سپس ورودی‌ها در هر لایه از طریق نورون‌ها و توابع تعریف شده به لایه خروجی انتقال می‌یابند و در این بین با یافتن برخی پارامترهای مجهول مانند بایاس و وزن هر نورون، مقادیر هدف را تخمین می‌زنند. سپس معیار ارزیابی شبکه برای محاسبه خطای بین مقادیر مشاهده شده و پیش‌بینی شده شبکه‌ها انتخاب می‌شود. وزن و بایاس، نورون‌ها را با الگوریتم‌های مختلف بر اساس داده‌های آموزشی تعیین می‌کند. این مرحله تا زمانی تکرار می‌شود که خطای بین مقادیر مشاهده شده و پیش‌بینی شده کمتر از یک آستانه شود. در نهایت مقادیر مناسبی برای وزن‌ها و بایاس‌ها یافت می‌شود، که دارای کمترین خطا بین مقادیر واقعی و پیش‌بینی است. در این مطالعه برای ارزیابی عملکرد گندم دیم از شبکه عصبی مصنوعی با ساختار سه لایه میانی در هر لایه حدود ۸ تا ۱۲ نورون قرار داشت استفاده شد. از الگوریتم لاونبرگ-مارکوارت^۱ برای آموزش ساختار شبکه عصبی استفاده شد. همچنین حداکثر تعداد تکرار ۱۰۰۰ عدد در نظر گرفته شد.



شکل ۳- نمونه از جنگل تصادفی ایجاد شده برای تحقیق حاضر

$$gcv(M) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2 / (1 - \frac{C(M)}{n})^2 \quad (3)$$

لامقدار وابسته پیش بینی شده توسط مدل، n تعداد مشاهدات در سری داده ها است، M تعداد عبارات غیر ثابت در مدل و C(M) تابع خطای مرکب است. هدف از C(M) جبران پیچیدگی مدل است، در این شرایط از برازش بیش از حد جلوگیری می شود و صرفه جویی در مدل افزایش می یابد. تابع C(M) به صورت $C(M) = M + Cd$ تعریف می شود که C عامل خطای هزینه است، که برای بهینه سازی تابع پایه توسط کاربر تعریف شده است و d درجه آزادی مؤثر است، که برابر با تعداد توابع پایه مستقل در مدل است. هر چه عامل C بیش تر باشد توابع بیش تری حذف می شوند. برای محاسبه اهمیت نسبی یک متغیر، تمام عباراتی را که دارای متغیرهای مورد سؤال هستند، حذف می شوند و مدل بار دیگر برازش داده می شود و سپس میزان کاهش برای برازش محاسبه می شود. هر متغیری که به هنگام حذف به کوچک شدن مدل کمک نماید، به عنوان مهم ترین متغیر شناخته می شود. متغیرهایی که نمرات کمتری دریافت می کنند از اهمیت کم تری برخوردارند. این نمرات مطابق با نسبت کاهش برازش مدل تولید شده است. اجرای مدل MARS در نرم افزار (MATLAB2018b) صورت گرفت. مدل MARS از تعدادی توابع پایه برای پیش بینی استفاده می کند، که در این تحقیق برای مدل سازی عملکرد گندم ۱۴ تابع پایه استفاده شد. جدول (۲) مقادیر مناسب پارامترهای موجود در MARS را برای این تحقیق بعد از سعی و خطاهای لازم بیان می کند

$$f(x) = \beta_0 + \sum_{m=1}^M \beta_m h_m(x) \quad (1)$$

با شروع از $h_0(x) = 1$ توابع پایه ای با شیوه ای تکرار شونده به مدل اضافه می شوند. برای هر $h_m(x)$ دو انتخاب وجود دارد:

$$(x - t)_+ = \begin{cases} x - t & \text{if } x > t, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

$$(t - x)_+ = \begin{cases} t - x & \text{if } t > x, \\ 0, & \text{otherwise,} \end{cases}$$

که در آن، x متغیر و مقدار ثابت t گره نامیده می شود. بر اساس هر متغیر، گره شناسایی شده (t) متفاوت خواهد بود و ضریب (β_m) در هر حالت $h_m(x - t)$ و $h_m(t - x)$ متغیر است. تعداد و مکان گره ها در روش MARS از طریق یک فرآیند گام به گام پیشرو و پسرو انجام می گیرد. در مرحله پیشرو، تعداد زیادی از گره ها تولید می شوند و در مرحله پسرو گره هایی که در برازش کلی مشارکت کم تری داشته اند، حذف می شوند. مدل MARS بهینه، در دو مرحله ساخته می شود. یک فرآیند انتخاب گام به گام پیشرو که با یک فرآیند بازگشتی دنبال می شود. انتخاب گام به گام پیشرو با استفاده از تابع پایه و یک مقدار ثابت آغاز می شود. در هر گام، از کلیه خطوط هر تابع پایه، فرآیند خطی انتخاب می شود، که نواقص مدل برازش یافته را به حداقل می رساند. توابع پایه که حداقل ممکن را به مدل می دهند، دارای گام های حذفی هستند و سپس مدل بهینه انتخاب می شود. نقصان شاخص برازش استفاده شده بر اساس معیار استاندارد شده مقابل است که به صورت رابطه (۳) تعریف می شود (Friedman., 1991):

که در آن C نشان دهنده ثابت مشخص مبادله بین خطای آموزشی و عبارت جریمه $w\|2$ است، و y_i نشان دهنده خروجی برآوردگر است، و ζ همراه با ζ' پارامترهای غیرمنفی هستند. برای رسیدگی به مسئله بهینه‌سازی معادله (۸)، ضریب‌های لاگرانژ معرفی شده‌اند، که بیان فرمول کمینه‌سازی را تسهیل می‌کند. علاوه بر این، یک تابع هسته برای تبدیل مسئله به یک مسئله رگرسیون غیرخطی معرفی شده است. توابع کرنل مختلفی مانند خطی، چندجمله‌ای، تابع پایه شعاعی^۱ (RBF) و هسته‌های سیگموئید برای استفاده در ساختارهای SVR معرفی شده‌اند. جدول (۳) مقادیر مناسب پارامترهای موجود در SVR را برای این تحقیق بیان می‌کند که در آن رگرسیون نوع دوم با مقدار c و nu به ترتیب برابر ۱۰ و ۰/۵۰ در نظر گرفته شده‌اند. کرنل نوع Radial Basis با مقدار گامای برابر ۰/۱۲۵ مناسب‌ترین نتیجه را نشان داد.

الگوریتم تجزیه مد تجربی گروهی

تجزیه مد تجربی گروهی (EEMD) یک روش تجزیه اطلاعات برای استخراج سیگنال‌ها در اطلاعات تولید شده توسط عملیات غیرخطی و غیرثابت را از داده‌های اصلی ارائه می‌دهد (Wu and Huang, 2009). در این روش، روش پیش‌پردازش داده‌ها به منظور تجزیه سیگنال‌های سری زمانی (داده‌های اصلی) به تعداد کم و محدود حالت‌های نوسانی، مطابق با مقیاس زمانی خاص محلی نسبت به روش تجزیه مد تجربی (EMD) توسعه می‌یابد (Rehman and Mandic, 2010). در واقع به دلیل مسائل و مشکلات مربوط به ترکیب (اختلاط) مد، وو و هوانگ روش تجزیه مد تجربی گروهی (EEMD) را معرفی نمودند (Wu and Huang, 2009). EEMD متشکل از (الک کردن) مجموعه‌ای از سیگنال‌های اضافه شده نویز سفید است. حالت‌های نوسانی در داده‌ها را می‌توان با استفاده از توابع مد ذاتی (IMFs) تعبیه شده، بیان کرد. تجزیه مد تجربی یک روش فرکانس زمانی خود تطبیقی است و با استفاده از IMF، سیگنالی را به عنوان مجموع حالت‌های ارتعاش آهسته و سریع با میانگین صفر نشان می‌دهد (Huang et al., 1998; Wu and Huang, 2009). در حالت کلی، یک تابع حالت ذاتی (IMF) یک حالت نوسانی ساده را در مقایسه با تابع هارمونیک ساده نشان می‌دهد. مراحل انجام این روش در شش مرحله به صورت خلاصه بیان شده است (Yeh et al., 2010):

مرحله اول: تشخیص تمام اکستریم‌ها، حاوی نقاط ماکزیمم و مینیمم محلی سری زمانی داده شده $(y(t))$:

به‌طوریکه تعداد درجات تعامل نهایی ۲ و مقدار معیار اعتبارسنجی تقاطعی تعمیم‌یافته هم ۰/۰۲۳۹ بدست آمدند. همچنین مقدار ضریب جریمه برای هر ۱۴ تابع پایه که در مدل نهایی دخیل بودند بین ۱ تا ۳ در نظر گرفته شدند. این ضرایب با کمینه‌سازی مجموع مربعات باقی‌مانده‌ها تعیین شدند. به طور کلی، زمانی که تفاوت بین عملکرد گندم دیم مشاهده شده و پیش‌بینی شده حداقل باشد، مقادیر پارامتر به عنوان مقادیر بهینه انتخاب شدند. همچنین بهینه‌ترین مدل MARS از مجموع ۱۴ تابع پایه با مقدار ثابت بدست آمد.

جدول ۲- پارامترهای بهینه انتخاب شده برای مدل MARS

Number of terms	۱۵
Number of basis functions	۱۴
Order of interactions	۱
Penalty	۲
Threshold	۰/۰۰۰۵
GCV error	۰/۰۲۳۹
Prune	بله

مدل رگرسیون بردار پشتیبان

رگرسیون بردار پشتیبان، یک ماشین یادگیری تحت نظارت است که به طور خاص برای وظایف رگرسیون طراحی شده است (Vapnik, 1998). این مدل از یک رویکرد برای به حداقل رساندن ریسک ساختاری استفاده می‌کند و از یک تابع هسته برای ایجاد یک ابر صفحه جدا کننده بهینه، استفاده می‌کند. این ابر صفحه با هدف به حداکثر رساندن حاشیه هندسی در حالیکه حد بالایی خطای تعمیم را به حداقل می‌رساند، توسط رابطه‌ی (۴) تعیین می‌گردد:

$$d = w^T x + b \quad (4)$$

که در آن w بردار وزن است، b بایاس است و d و x متعلق به نمونه $J = \{x_i, d_i\}_{i=1}^N$ تمرینی هستند. با در نظر گرفتن تابع ضرر غیر حساس به ε ، ضرایب y و d با به حداقل رساندن تابع ریسک به دست می‌آیند:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^N |y_i - d_i|_{\varepsilon} \quad (5)$$

که در معرض محدودیت‌های آورده شده در معادلات (۶) تا (۹) است.

$$d_i - y_i \leq \varepsilon + \zeta_i \quad (6)$$

$$y_i - d_i \leq \varepsilon + \zeta'_i \quad (7)$$

$$0 \leq \zeta_i \quad (8)$$

$$0 \leq \zeta'_i \quad (9)$$

جدول ۳- مقادیر مناسب برای پارامترها و توابع در نظر گرفته شده مدل SVR

متغیر	مقادیر
Number of independents	۸
SVM type	Regression type ۲ (C=۱۰/۰, nu=۰/۵۰)
Kernel type	Kernel type= Radial Basis Function (gamma=۰/۱۲۵)
Number of SVs	۱۳۰ (۱۰۲ bounded)

محل‌ی و پوشی که به وسیله‌ی مینیمم محلی به وجود می‌آید، صفر است. در این تحقیق برای پارامترهای ورودی از ۷ تابع مد ذاتی و ۱ باقیمانده استخراج شده استفاده شد. در نتیجه در ترکیب EEMD با مدل‌های یادگیری ماشین همواره دقت پیش‌بینی بهبود می‌یابد.

معیارهای ارزیابی خطا

به منظور ارزیابی نتایج حاصل از اجرای هر یک از مدل‌ها و الگوهای ورودی مختلف از شش شاخص شامل، ضریب همبستگی^۱ (CC)، جذر میانگین مربعات خطا^۲ (RMSE)، میانگین خطای مطلق^۳ (MAE)، ضریب تبیین^۴ (R²) و شاخص نش ساتکلیف^۵ (NSE) استفاده گردید. همچنین از شاخص پراکندگی^۶ (SI) برای ارزیابی اثربخشی عملکرد مدل‌های محاسباتی نرم در دو مرحله توسعه (یعنی مراحل آموزش و آزمون) مربوط به ارزیابی عملکرد دیم گندم استفاده شد.

در این روابط x_0 و x_p به ترتیب i امین داده شبیه‌سازی شده توسط مدل و مقدار واقعی (مشاهده شده)، \bar{x}_0 و \bar{x}_p به ترتیب میانگین کل داده‌های x_0 و x_p در جامعه آماری و n تعداد نمونه‌ها است.

نتایج و بحث

شکل (۴) مقادیر ضریب همبستگی بین متغیرهای ورودی نسبت به یکدیگر و نسبت به مقادیر عملکرد گندم دیم را نشان می‌دهد. بدین صورت که محورهای افقی و عمودی متغیرهایی که بر عملکرد گندم تأثیر می‌گذارند قرار دارند. سپس ضریب همبستگی بین آن‌ها محاسبه شده است. میزان همبستگی به صورت طیف رنگی مشخص شده است. همان‌طور که مشاهده می‌گردد بیش‌ترین ضریب همبستگی نسبت به مقادیر عملکرد گندم مربوط به متغیرهای R ، T_{\max} ، T_{\min} و T_{amin} (میانگین حداقل دمای مطلق فصل زراعی)

مرحله دوم: نقاط ماکزیمم محلی به دست آمده را به منظور ایجاد یک $e_{\max}(t)$ روی پوش بالا و همچنین تمام نقاط حداقل برای به دست آوردن $e_{\min}(t)$ روی پوش پایین‌تر به ترتیب با درون‌یابی اسپلاین محاسبه می‌شود،
مرحله سوم: میانگین $m(t)$ دو پوش محاسبه شده با استفاده از رابطه زیر محاسبه می‌گردد:

$$m(t) = \frac{e_{\max}(t) + e_{\min}(t)}{2} \quad (10)$$

مرحله چهارم: رابطه‌ی (۱۱) میانگین دو پوش محاسبه شده را از داده‌های استخراج شده (IMF) کاهش می‌دهد:

$$h(t) = y(t) - m(t) \quad (11)$$

مرحله پنجم: اگر $h(t)$ معیار توقف از پیش تعریف شده برقرار باشد، $h(t)$ به عنوان اولین IMF تعریف می‌شود که به صورت $C_1(t)$ نوشته می‌شود و اولین مرحله آن است. ولی اگر $h(t)$ معیار توقف برقرار نباشد $y(t)$ و $h(t)$ دوباره محاسبه می‌شود و مراحل ۱ تا ۴ دوباره تکرار می‌شود تا زمانی که معیار توقف برقرار باشد و همچنین $h(t)$ شروط IMF را برقرار کند.

مرحله ششم: سپس اولین باقیمانده $r_1(t) = y(t) - C_1(t)$ به عنوان داده‌های جدیدی در نظر گرفته می‌شود، در مرحله آخر زمانی که باقیمانده به یک روند یکنواخت تبدیل شود یا یک نقطه ماکزیمم و مینیمم محلی را شامل شود، روند غربال‌گری متوقف می‌شود، به طوریکه دیگر نمی‌توان IMF را استخراج کرد (Huang et al., 2003). سیگنال اصلی را می‌توان در پایان این روند غربال‌گری با استفاده از مجموع IMF و باقیمانده با معادله (۱۲) محاسبه کرد:

$$y(t) = \sum_{j=1}^n c_j(t) + r_n \quad (12)$$

در این معادله $r_n(t)$ مقدار باقیمانده نهایی، n تعداد IMF های استخراج شده و $c_i(t)$ مدهای ذاتی استخراج شده را نشان می‌دهد. آخرین IMF استخراج شده، کم‌ترین مؤلفه فرکانسی است، که برای انجام آن، یک IMF باید دو شرط را ارضا کند (Huang et al., 1998)

(۱) تعداد اکسترمم N_e (ماکزیمم و مینیمم محلی) و تعداد عبور از صفرها N_z باید برابر باشند یا حداکثر یکی اختلاف داشته باشند.

$$\leq (N_z - 1) \leq N_e (N_z + 1) \quad (13)$$

(۲) در هر نقطه، مقدار میانگین پوشی که به وسیله‌ی ماکزیمم

- 1- Correlation Coefficient
- 2- Root Mean Square Error
- 3- Mean Absolute Error
- 4- R- squared statistic (Coefficient of Determination; Pearson's r squared)
- 5- Nash-Sutcliffe model efficiency coefficient
- 6- Scatter Index

است، که مقادیر ضریب همبستگی بیشتر از (+۰/۶۰) را دارند.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_o - x_p)^2}{n}} \quad 0 \leq RMSE \leq +\infty \quad (14)$$

$$NSE = 1 - \frac{\sum_{i=1}^n (x_o - x_p)^2}{\sum_{i=1}^n (x_o - \bar{x}_o)^2} \quad -1 \leq NSE \leq 1 \quad (15)$$

$$CC = \frac{\sum_{i=1}^N (x_p - \bar{x}_p)(x_o - \bar{x}_o)}{\sqrt{\sum_{i=1}^N (x_p - \bar{x}_p)^2 \sum_{i=1}^N (x_o - \bar{x}_o)^2}} \quad 0 \leq R \leq 1 \quad (16)$$

$$MAE = \frac{\sum_{i=1}^n |x_o - x_p|}{n} \quad 0 \leq MAE \leq +\infty \quad (17)$$

$$SI = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^N [(x_p - \bar{x}_p) - (x_o - \bar{x}_o)]^2}}{\bar{x}_o} \quad 0 \leq SI \leq +\infty \quad (18)$$

$$R^2 = \left[\frac{\sum_{i=1}^N (x_p - \bar{x}_p)(x_o - \bar{x}_o)}{\sqrt{\sum_{i=1}^N (x_p - \bar{x}_p)^2 \sum_{i=1}^N (x_o - \bar{x}_o)^2}} \right]^2 \quad 0 \leq R^2 \leq 1 \quad (19)$$

های آموزش و معیارهای خطای، $RMSE=0/140(\text{ton.ha}^{-1})$ ، $MAE=0/110(\text{ton.ha}^{-1})$ و $SI=0/128$ ، $NSE=0/914$ برای داده‌های آزمون نسبت به سه مدل دیگر از دقت بالایی برخوردار بوده است، مدل‌های SVM و MLP با معیارهای خطای، $RMSE=0/155(\text{ton.ha}^{-1})$ ، $MAE=0/114(\text{ton.ha}^{-1})$ ، $SI=0/14$ و $NSE=0/894$ برای داده‌های آزمون نتایج نزدیکی به مدل منتخب منفرد داشته‌اند. نکته‌ای که حائز اهمیت است، دقت بالای مدل‌ها در برآورد داده‌های آموزش نسبت به داده‌های آزمون است، به عبارتی، مدل‌ها برای پیش‌بینی بهتر، نیاز به داده‌های بیش‌تری داشته‌اند. در شبیه‌سازی داده‌های آموزش، مدل EEMD-ANN با معیارهای خطای، $RMSE=0/099(\text{ton.ha}^{-1})$ ، $MAE=0/073(\text{ton.ha}^{-1})$ و $SI=0/088$ بیش‌ترین دقت را داشته است. همچنین عملکرد این مدل در شبیه‌سازی داده‌های آزمون نیز با معیارهای خطای، $RMSE=0/121(\text{ton.ha}^{-1})$ ، $MAE=0/087(\text{ton.ha}^{-1})$ و $SI=0/110$ ، $NSE=0/936$ بهتر بوده است. در شبیه‌سازی داده‌های آموزش، دیگر مدل ترکیبی EEMD-SVR با معیارهای خطای، $RMSE=0/122(\text{ton.ha}^{-1})$ ، $MAE=0/087(\text{ton.ha}^{-1})$ و $SI=0/109$ ، $NSE=0/922$ در اولویت بعدی قرار دارد.

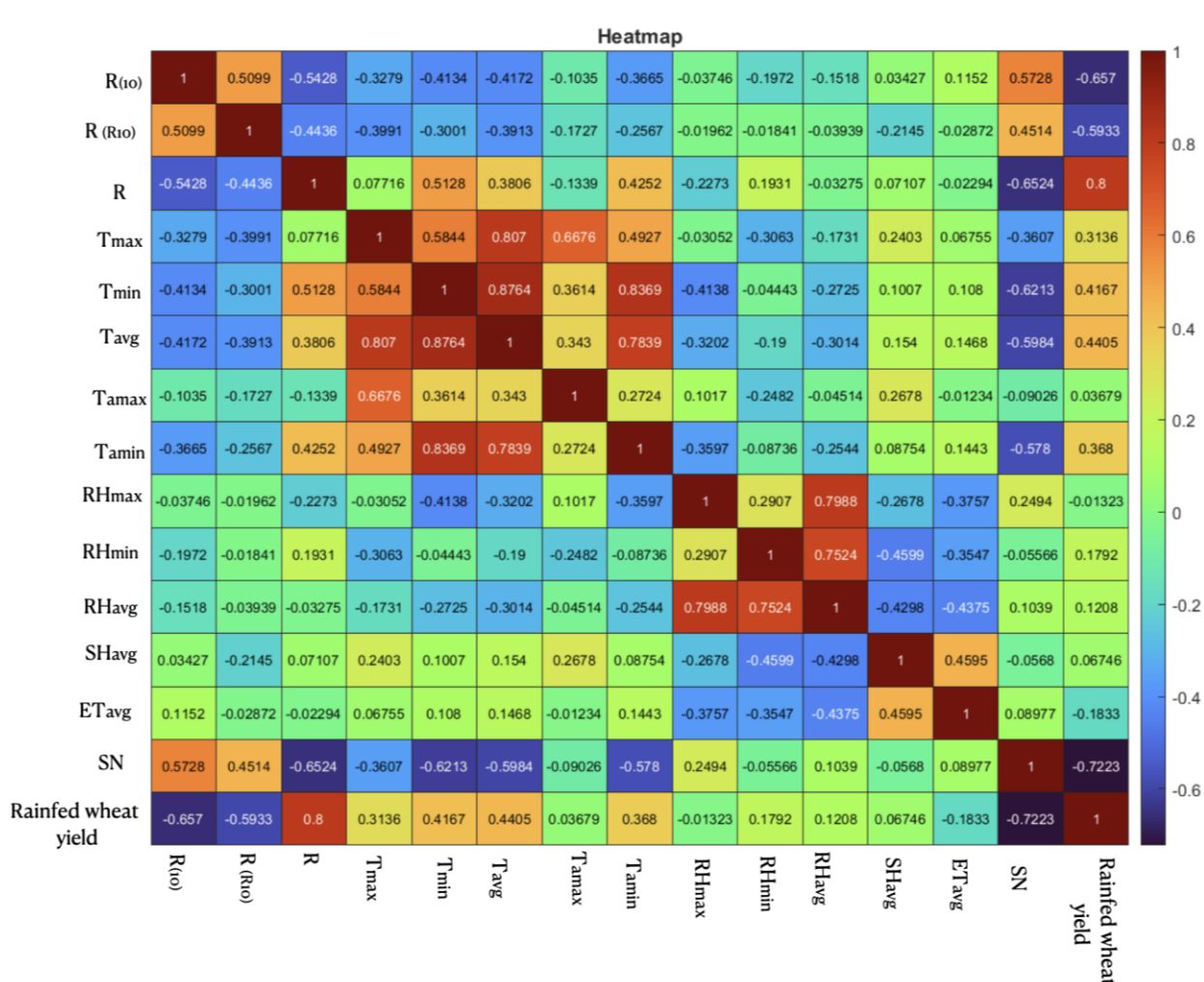
اما در شبیه‌سازی داده‌های آزمون، که اساس انتخاب مدل برتر نیز است، مدل EEMD-MARS با عملکردی نزدیک به مدل منتخب داده‌های آموزش EEMD-ANN، دقت بیش‌تری داشته است، به طوری که در این مدل، معیارهای خطای برای داده‌های آزمون برابر $RMSE=0/112(\text{ton.ha}^{-1})$ و $MAE=0/088(\text{ton.ha}^{-1})$ به دست آمد. شایان ذکر است که مدل RF در برآورد داده‌های آموزش با معیارهای خطای، $RMSE=0/190(\text{ton.ha}^{-1})$ ، $MAE=0/811$ ، $SI=0/169$ و $NSE=0/879$ در برآورد مرحله آزمون با

این نشان می‌دهد که مقدار عملکرد گندم با مقادیر بارش و دما همبستگی بالایی دارد که با افزایش بارندگی بر میزان عملکرد گندم دیم تاثیر مثبت مؤثری دارد. در حالیکه عملکرد گندم دیم نسبت به شمار اولین بارش مؤثر ۱۰ میلی‌متری، شمار آخرین بارش مؤثر ۱۰ میلی‌متری (R_{R10}) به همراه مجموع تعداد روزهای یخبندان (SN) همبستگی نسبتاً قوی (در محدوده ۰/۵ تا ۰/۷۵) معکوس دارد، که با افزایش متغیرهای مذکور عملکرد گندم دیم با کاهش همراه است. با این نتایج ضریب همبستگی نشان می‌دهد، که متغیرهایی مانند دما، نسبت به سایر پارامترها با توجه به مقادیر ضریب همبستگی از نقش مؤثرتری برخوردار هستند. از طرفی ضریب همبستگی دمای حداکثر مطلق با میزان عملکرد گندم دیم (۰/۳۶۷) است که بسیار ناچیز است، در صورتیکه سایر پارامترهای دما از ضرایب همبستگی بالاتری برخوردار هستند. در نتیجه با توجه به اطلاعات موجود در این شکل بینش مناسبی از نحوه ترکیب متغیرهای ورودی به مدل‌ها جهت تخمین عملکرد گندم دیم بیان می‌کند.

نتایج ارزیابی مدل‌های پیش‌بینی کننده عملکرد گندم دیم برای داده‌های آموزش و آزمون در جدول (۴) ارائه شده است. در مرحله اول، مطابق نتایج به دست آمده از دو گروه داده‌های آزمون و آموزش، این استنباط وجود دارد که مدل‌های مورد استفاده عملکرد قابل قبولی در تخمین عملکرد محصول داشته‌اند. مقایسه بین مدل‌های منفرد با ترکیبی نشان می‌دهد که مدل‌های ترکیبی در حالت کلی از دقت شبیه‌سازی بالایی برخوردار بوده‌اند، به عبارتی می‌توان گفت، الگوریتم‌های آموزش دهنده در مدل‌های ترکیبی با استفاده از یک الگوریتم دیگر تقویت شده و داده‌های خروجی را با دقت بیش‌تری مدل‌سازی کرده است. در بین مدل‌های منفرد، مدل MARS با معیارهای خطای، $RMSE=0/152(\text{ton.ha}^{-1})$ ، $MAE=0/114(\text{ton.ha}^{-1})$ و $SI=0/136$ ، $NSE=0/879$ برای داده-

مدل‌های دیگر عملکرد ضعیف‌تری داشته است.

معیارهای خط‌مشی، $RMSE=0/193(\text{ton}\cdot\text{ha}^{-1})$ ، $MAE=0/119(\text{ton}\cdot\text{ha}^{-1})$ نسبت به $SI=0/176$ و $NSE=0/836$



شکل ۴- مقادیر ضریب همبستگی بین متغیرهای ورودی و خروجی عملکرد گندم دیم

برای انتخاب ویژگی و کاهش ابعاد پارامترهای ورودی اثرگذار بر خروجی مسئله، یعنی پیش‌بینی عملکرد گندم دیم استفاده گردید. در الگوریتم ژنتیک مقادیر جمعیت اولیه، نرخ جمعیت و نرخ جهش به ترتیب ۱۰۰، ۰/۵ و ۰/۰۵ در نظر گرفته شد. مقدار کم ضریب گاما، نشان‌دهنده یک انتخاب خوب از زیرمجموعه‌های پیش‌بینی و همچنین یک مدل دقیق برای پیش‌بینی است. بر این اساس، نتایج به دست آمده برای انتخاب بهترین پارامترهای ورودی به مدل‌های پیشنهادی، در جدول (۵) آورده شده است. برای این منظور در این تحقیق ۱۰۰ سناریو (ماسک) در نظر گرفته شد، و از بین آن‌ها سناریو مناسب که دارای کمترین گاما و کمترین خطای استاندارد و همچنین کمترین شاخص بود انتخاب شد. نتایج در جدول (۵) آورده شده است. طبق این جدول، شاخص (V_{ratio}) و آماره گاما (GT) کمترین مقدار

همچنین بیشترین ضریب تبیین در بین مدل‌های منفرد برای داده‌های آموزش با ۰/۸۸۵ مربوط به مدل ANN و برای داده‌های آزمون با ۰/۹۲۱ مربوط به مدل MARS است. ترکیب مدل‌های یادگیری ماشین با الگوریتم تجزیه مد تجزیه گروهی باعث افزایش دقت مدل و همچنین افزایش ضریب تبیین شده است، در بین مدل‌های ترکیبی مدل EEMD-MARS برای داده‌های آزمون با ضریب تبیین ۰/۹۵۲ از دقت بالایی نسبت به سایر مدل‌های ترکیبی دیگر برخوردار بوده است، در نتیجه، مدل EEMD-MARS شبیه‌سازی دقیق‌تری انجام داده است.

نتایج آزمون گاما برای انتخاب تعداد بهینه متغیرهای ورودی در این تحقیق، از روش ترکیبی آزمون گاما با الگوریتم ژنتیک

مجموع بارش (R)، میانگین حداکثر دما (T_{max})، میانگین حداقل دما (T_{min})، میانگین متوسط دما (T_{avg})، میانگین حداکثر رطوبت نسبی (RH_{max})، میانگین متوسط رطوبت نسبی (RH_{avg}) و تعداد روزهای یخبندان (SN) می‌باشد.

ممکن در بین ترکیبات معنادار را دارد و نزدیک به صفر می‌باشند. با حداقل کردن تابع هدف در الگوریتم ژنتیک، بهینه‌ترین ترکیب ورودی جهت پیش‌بینی عملکرد گندم دیم به دست آمد. بر این اساس بهترین ترکیب ورودی با استفاده از ترکیب آزمون گاما با الگوریتم ژنتیک شامل پارامترهای شمار روز اولین بارش مؤثر ۱۰ میلی‌متری (R_{10}),

جدول ۴- نتایج معیارهای ارزیابی خطای مدل‌های پیش‌بینی کننده عملکرد گندم به تفکیک داده‌های آموزش و آزمون

مدل	آموزش					آزمون				
	RMSE (ton.ha ⁻¹)	MAE (ton.ha ⁻¹)	SI	NSE	R ²	RMSE (ton.ha ⁻¹)	MAE (ton.ha ⁻¹)	SI	NSE	R ²
MLP	0/148	0/107	0/132	0/885	·/885	0/155	0/114	0/140	0/894	·/909
EEMD-ANN	0/099	0/073	0/088	0/949	·/941	0/121	0/087	0/110	0/936	·/938
SVM	0/161	0/118	0/144	0/864	·/865	0/155	0/114	0/142	0/894	·/912
EEMD-SVR	0/122	0/087	0/109	0/922	·/923	0/132	0/080	0/121	0/923	·/929
RF	0/190	0/121	0/169	0/811	·/823	0/193	0/119	0/176	0/836	·/878
EEMD-RF	0/146	0/110	0/130	0/888	·/905	0/156	0/115	0/141	0/893	·/929
MARS	0/152	0/114	0/136	0/879	·/879	0/140	0/110	0/128	0/914	·/921
EEMD-MARS	0/134	0/093	0/119	0/906	·/907	0/112	0/088	0/101	0/945	·/952

جدول ۵- مقادیر به دست آمده از آزمون گاما با الگوریتم ژنتیک برای انتخاب بهترین پارامترهای بهینه

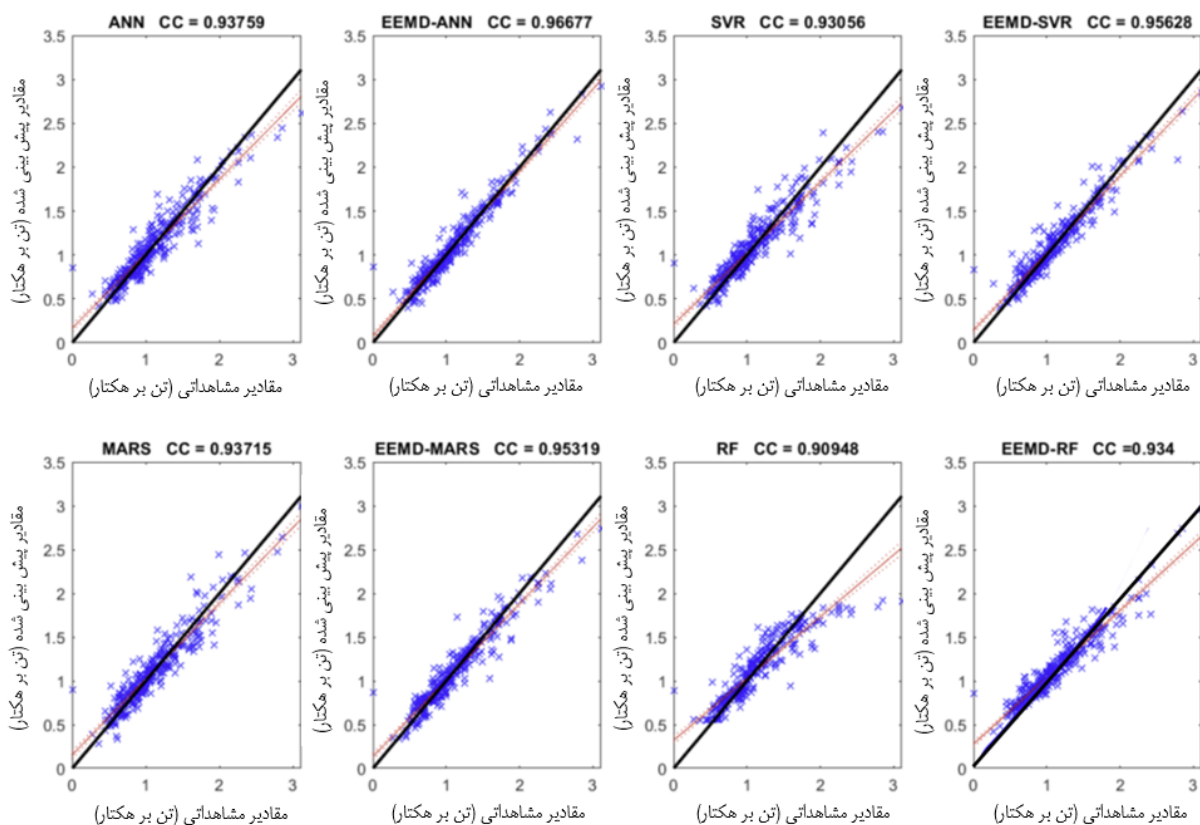
MASK	گاما	خطای استاندارد	گرادیان	Vratio (شاخص)
{1·01111·00101001}*	·/۰۴۸۵۰۳	·/۰۰۶۷۳۵۵	$۵/۱۰۱ \times 10^{-8}$	·/۲۴۱۸۶

* {R (مجموع بارش فصل زراعی), T_{max} (میانگین حداکثر دمای فصل زراعی), T_{min} (میانگین حداقل دمای فصل زراعی), T_{avg} (میانگین متوسط دمای فصل زراعی), RH_{max} (میانگین حداکثر رطوبت نسبی فصل زراعی), RH_{avg} (میانگین متوسط رطوبت نسبی فصل زراعی), SN (تعداد روزهای یخبندان), R_{10} (شمار روز اولین بارش مؤثر ۱۰ میلی‌متری)}

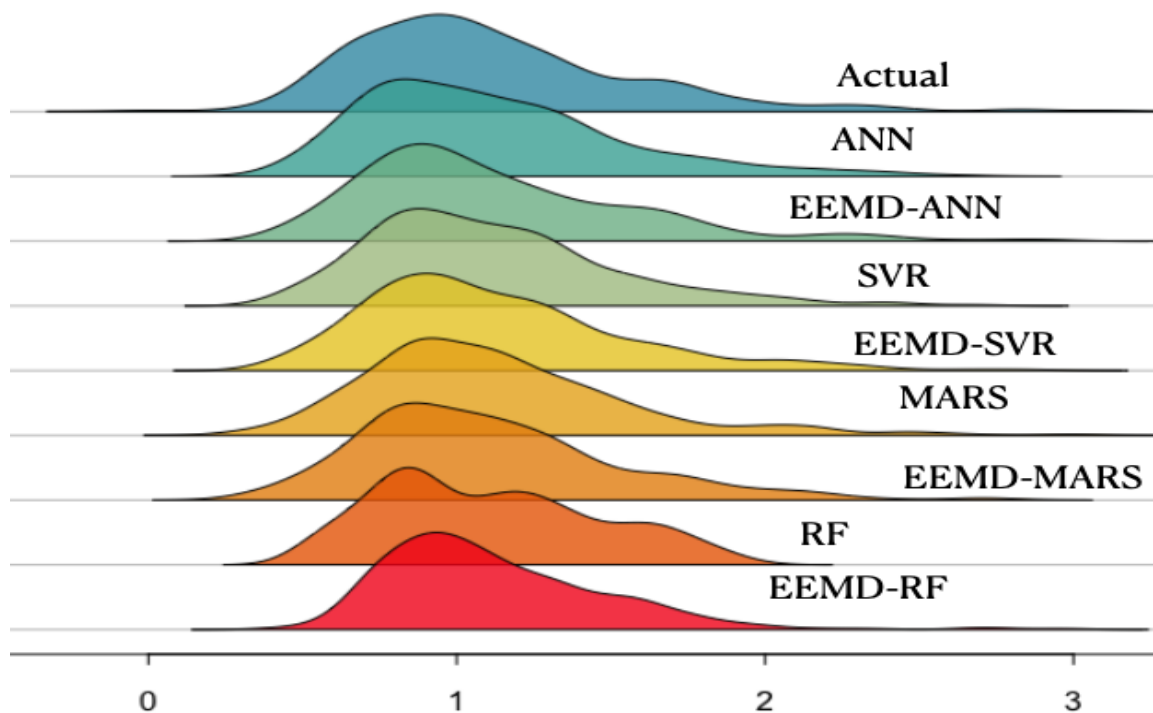
گندم دیمی داشتند.

در ادامه مقایسه مدل‌ها، از نمودار (Ridgeline) برای تجسم توزیع داده‌های شبیه‌سازی شده با مدل‌های یادگیری ماشینی و داده‌های مشاهده‌ای استفاده شد، و نتایج آن در شکل (۶) ارائه شده است. مدل‌ها همواره در برآورد مقادیر کمینه، مقادیر را نسبت به مقادیر مشاهده‌ای به مقدار کمتری برآورد کرده‌اند. که اختلاف خطای این مقدار در مدل RF نسبت به سایر مدل‌ها بیشتر است. همچنین مدل RF، داده‌های با تراکم بیشتر را متفاوت از داده‌های مشاهده‌ای و دیگر مدل‌ها برآورد کرده است. در محدوده داده‌های بین ۰/۵ تا ۲ تن بر هکتار، مدل‌ها تقریباً داده‌ها را مطابق با داده‌های مشاهده‌ای تخمین زده‌اند. در برآورد داده‌های بیشینه و بیشتر از ۲ تن بر هکتار، مدل‌های ترکیبی تا حد زیادی این مقادیر را متناسب با داده‌های مشاهده‌ای برآورد کرده‌اند، اما مدل‌های منفرد، بخصوص مدل RF این مقادیر را با مقدار کمتری برآورد کرده است. همچنین شکل توزیع داده‌های مدل‌های ترکیبی به شکل توزیع داده‌های مشاهده‌ای نزدیک‌تر و همخوانی بیشتری دارند.

شکل (۵) مقادیر ضریب همبستگی مقادیر مشاهداتی و پیش‌بینی شده توسط مدل‌ها، برای کل داده‌های مورد استفاده را نشان می‌دهد، که در آن محور افقی بیان کننده مقادیر مشاهداتی است و مقادیر محور عمودی نیز بیانگر مقادیر پیش‌بینی عملکرد گندم دیم می‌باشد. همان‌طور که از نمودارهای موجود مشخص است، ضریب همبستگی همه مدل‌های به کار رفته همواره بیش‌تر از ۰/۹۰ هستند. پراکندگی داده‌ها در مدل‌های هیبریدی کم‌تر از مدل‌های منفرد مشاهده می‌گردد. همچنین در هر مدل به کار گرفته شده، مقادیر ضریب همبستگی مدل‌های هیبریدی مانند EEMD-MARS و EEMD-RF به ترتیب با ضریب همبستگی ۰/۹۵۳ و ۰/۹۳۴ بیشتر از مدل‌های پایه خودشان با ضرایب همبستگی $MARS=0/۹۳۷$ و $RF=0/۹۰۹$ هستند. در بین مدل‌های مورد استفاده کمترین ضریب همبستگی، مربوط به مدل منفرد جنگل تصادفی با مقدار حدود ۰/۹۰۹ است. در حالیکه مدل EEMD-MARS دارای بالاترین ضریب همبستگی بود. این نتایج تا حدودی با نتایج جدول (۴) همخوانی دارد. علاوه بر این، نتایج کلی نشان می‌دهد که مدل‌ها بخصوص مدل‌های هیبریدی عملکرد مناسبی در تخمین عملکرد



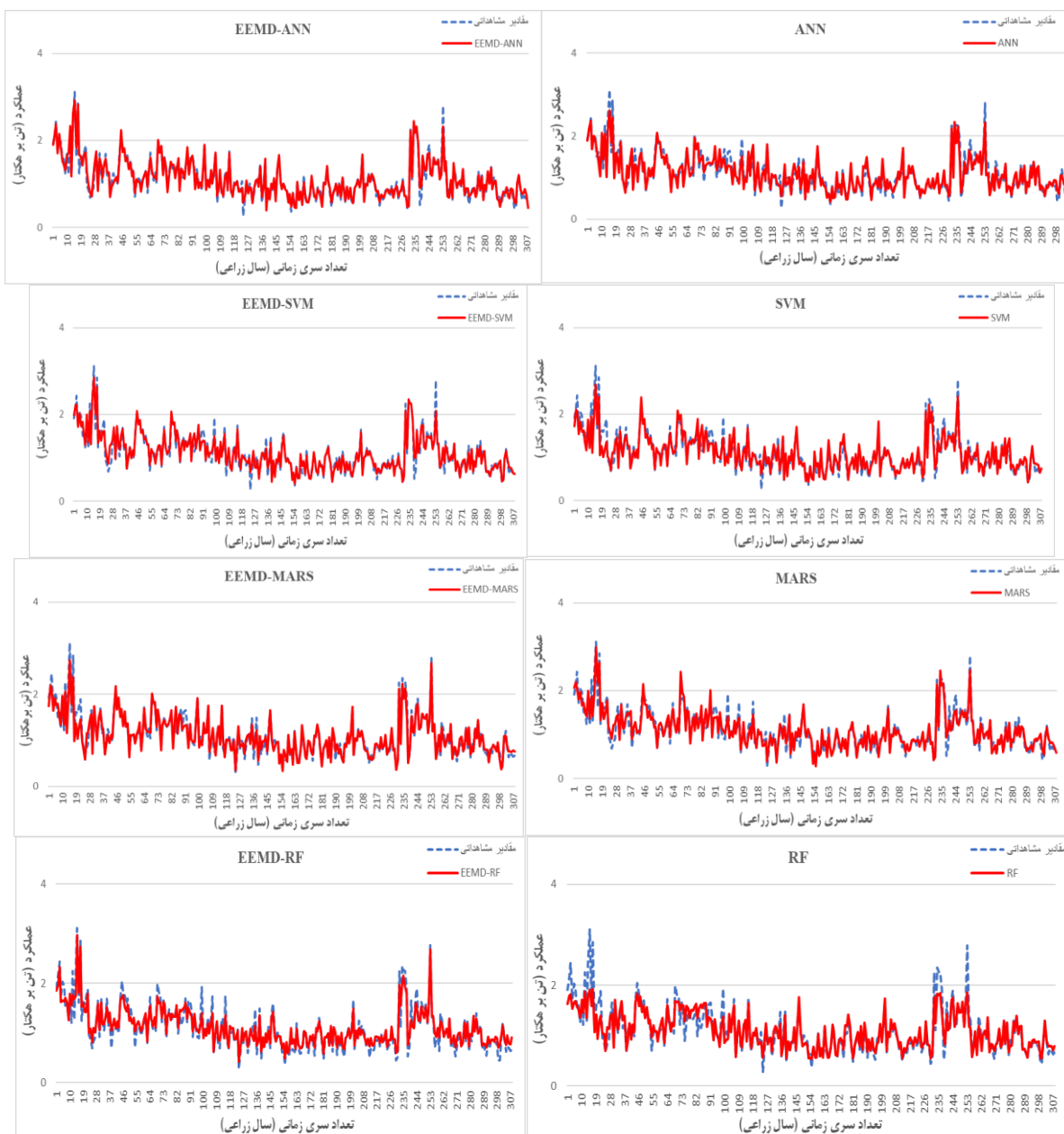
شکل ۵- مقادیر همبستگی داده‌های مشاهداتی و پیش‌بینی شده با مدل‌های یادگیری ماشین، (ضریب همبستگی: CC)



شکل ۶- دیاگرام Ridgeline داده‌های مشاهداتی و شبیه‌سازی شده

آن‌ها مقادیر حدی تا حدودی مناسب‌تر تخمین زده شده است. در نتیجه می‌توان چنین بیان کرد که مدل‌های هیبریدی EEMD-ANN، EEMD-SVR، EEMD-MARS و EEMD-RF عملکرد مناسبی در تخمین مقادیر حداقل و حداکثر عملکرد گندم را داشتند، که برتری نسبی نسبت به مدل‌های یادگیری ماشین منفرد داشتند.

در نهایت، مقادیر عملکرد گندم توسط مدل‌های مورد استفاده تخمین زده شد، که نتایج آن‌ها به صورت نمودار خطی مانند آنچه که در شکل (۷) قابل مشاهده است، بیان شده است. نتایج نشان می‌دهد، که همه مدل‌ها مقادیر پایه (کمتر از میانگین کل داده‌ها) را با دقت مناسبی پیش‌بینی کرده‌اند. مقادیر حدی توسط مدل‌های یادگیری ماشین همواره با دقت کمتری انجام شده است، اما در مدل هیبریدی



شکل ۷- نمودار سری زمانی داده‌های مشاهداتی و شبیه‌سازی شده

پردازش داده‌های ورودی، به منظور انتخاب ترکیبی مناسب از آن‌ها در

یکی از مراحل مهم و پیچیده برای مدل‌سازی غیرخطی، پیش-

است که پیش‌تر در تحقیقات امیرعشایری و همکاران در استفاده از آن برای بهبود دقت تبخیر-تعرق مرجع روزانه توسط مدل‌های درخت تصمیم و شبکه عصبی مصنوعی با روش پیش‌پردازش‌کننده گروهی یافت شد (Amirashayeri et al., 2023). لذا با نتایج این تحقیق همسو می‌باشد. در نتیجه می‌توان آن را برای استفاده در سایر زمینه‌ها نیز توصیه نمود. در نتیجه استفاده از سایر مدل‌های یادگیری ماشین در پیش‌بینی میزان عملکرد گندم دیم برای رسیدن به نتایج مورد انتظار و مقایسه با نتایج تحقیق حاضر همواره به محققان توصیه می‌گردد.

نتیجه‌گیری

ایران به دلیل قرار گرفتن در ناحیه خشک و نیمه‌خشک همواره با کمبود بارندگی و تنش منابع آب همراه است. با بحرانی شدن بیشتر آبخوان کشور، توسعه سطح زیرکشت و بهبود عملکرد گندم دیم، یکی از راه‌کارهای کم هزینه و پایدار برای مدیریت تأمین آب در این شرایط می‌باشد. از طرفی مدل‌های یادگیری ماشین یکی از ابزارهای مناسب در جهت شبیه‌سازی در این زمینه هستند، که با کم‌ترین داده‌های در دسترس معمولاً با دقت مناسبی همراه هستند. لذا در این تحقیق از مدل‌های یادگیری ماشین برای بررسی عملکرد گندم دیم در شهرستان‌های استان آذربایجان غربی استفاده شد. طیف وسیعی از مدل‌های مختلفی که بر مبنای رگرسیون (ANN، MARS، SVR) و درخت تصمیم (RF) هستند برای این منظور انتخاب شدند، که عملکرد آن‌ها توسط برخی معیارهای ارزیابی خطا نظیر RMSE، MAE، SI و نمودارهای بصری مورد ارزیابی قرار گرفت. ۱۴ پارامتر هواشناسی به عنوان پارامتر ورودی به مدل‌ها جمع‌آوری شدند که بر میزان عملکرد گندم می‌توانند تأثیر بگذارند. با استفاده از آزمون گاما تست و در ترکیب با الگوریتم ژنتیک هشت پارامتر بهینه انتخاب شدند، که شامل پارامترهای مجموع بارش فصل زراعی، شمار روز اولین بارش مؤثر ۱۰ میلی‌متری، میانگین حداکثر دمای فصل زراعی، میانگین حداقل دمای فصل زراعی، میانگین متوسط دمای فصل زراعی، میانگین حداکثر رطوبت نسبی فصل زراعی، میانگین رطوبت نسبی فصل زراعی و تعداد روزهای یخبندان فصل زراعی بودند. نتایج بیانگر عملکرد مناسب مدل‌ها بود. با توجه به مدل‌های مذکور مدل MARS بهینه با ۱۴ تابع رگرسیونی و معیار اعتبار سنتی تعمیم‌یافته 0.28135 و مدل توسعه داده شده آن توسط EEMD با ۷ تابع مد ذاتی و ۱ باقیمانده برای پارامترهای مستقل مناسب‌ترین عملکرد را در تخمین عملکرد گندم دیم داشتند. همچنین استفاده از الگوریتم EEMD جهت بهبود دقت پیش‌بینی در ترکیب با مدل‌ها موجب افزایش قابل توجه دقت مدل‌ها گردید، که می‌تواند یک گام مؤثر در زمینه پیش‌بینی دقت چنین پارامترهایی باشد. شیب

مدل می‌باشد. انتخاب متغیرهای ورودی مناسب برای مدل‌های هوشمند مصنوعی و روش‌های داده‌کاوی بسیار در عملکرد هر یک از مدل‌ها اثرگذار می‌باشد. از این‌رو، بر اساس مشاهدات تجربی و مطالعات سایر محققان از جمله متغیرهای ورودی مؤثر در پیش‌بینی عملکرد گندم دیم می‌توان پارامترهای هیدرولوژیک نظیر بارش، شمار روز اولین و آخرین بارش مؤثر ۱۰ میلی‌متری، دما، تبخیر، رطوبت نسبی، تابش خورشید و تعداد روزهای یخبندان اثرگذاری قابل توجهی بر خروجی مدل‌ها دارند، استفاده نمود. نتایج آزمون گاما نشان داد که در صورت وجود تعداد زیادی از متغیرهای ورودی می‌توان متغیرهای ورودی که دارای بیشترین تأثیر در پیش‌بینی هستند را انتخاب نمود. این عمل می‌تواند موجب صرفه‌جویی در زمان و هزینه در مدل‌سازی‌ها گردد، که استفاده از آن برای سایر زمینه‌ها توصیه می‌گردد. همچنین ترکیب الگوریتم ژنتیک با آزمون گاما تست نتایج مطلوبی به همراه داشت که نتایج آن همسو با تحقیقات خوبان و همکاران در استفاده از آن برای انتخاب پارامترهای ورودی مناسب برای پیش‌بینی ترافیک جاده‌ای بود. نتایج آن نشان داد که بسیاری از پارامترهایی که در حالت طبیعی ممکن است تأثیر بسزایی در ترافیک جاده‌ای داشته باشند، در استفاده از مدل‌های یادگیری ماشین در نظر گرفته نمی‌شوند (Khouban et al., 2015). بنابراین می‌توان چنین بیان کرد که مدل‌های یادگیری ماشین بر مبنای داده‌هایی هستند که نسبت به زمان یا مکان متغیر باشند. نتایج مدل‌های یادگیری ماشین نشان داد که همواره این مدل‌ها ابزاری کارآمد در جهت پیش‌بینی عملکرد گندم دیمی هستند. مشابه این نتایج پیش‌تر توسط جعفری و همکاران (۱۴۰۲) به دست آمده بود، که در آن از شبکه عصبی مصنوعی و انواع الگوریتم‌های آموزش شبکه عصبی مصنوعی در برآورد عملکرد گندم دیم استان کرمانشاه استفاده کرده بودند. تنوع مدل‌ها در این تحقیق نشان داد که دقت آن‌ها نسبت به یکدیگر کمی می‌تواند متفاوت باشد، که نیازمند بررسی بیشتر می‌باشد که در زمینه‌های پیش‌بینی، پیش‌تر توسط سایر محققین پیشنهاد شده بود (Kayhomayoon et al., 2022). در بین مدل‌ها، عملکرد MARS نشان داد که می‌تواند با دقت مناسبی مقدار عملکرد گندم دیم را تخمین نماید. در حالت کلی نیز MARS همواره نشان داده است که یک مدل مناسب برای پیش‌بینی سری‌های زمانی و مکانی است. این نتیجه در تحقیقات رضایی بالف و همکاران در استفاده از MARS برای پیش‌بینی جریان رودخانه نیز دیده شد (Rezaie-Balfet et al., 2019). همچنین جمنانی و همکاران نیز به طور مشابه به چنین نتیجه‌ای در پیش‌بینی جریان ورودی به سد توسط مدل MARS رسیده بودند (Jamnani et al., 2024). عملکرد مدل‌های هیبریدی و نقش EEMD در بهبود دقت پیش‌بینی در این تحقیق قابل توجه بود که در نتایج آن در تحلیل معیارهای ارزیابی خطا و نمودارهای مورد استفاده بررسی شد. لذا این یک گام مؤثر در بهبود دقت پیش‌بینی

موردی: ایستگاه کرمان)، نشریه آب و خاک. ۲۴(۶): ۱۱۹۶-۱۱۸۶.

فرج زاده اصل، م. و زرین، آ. ۱۳۸۱. مدلسازی میزان عملکرد محصول گندم دیم با توجه به معیارهای اقلیم‌شناسی کشاورزی در استان آذربایجان غربی. برنامه‌ریزی و آمایش فضا. ۲۴(۲): ۷۷-۹۷.

قباتی سوق، م.، مساعدی، ا.، حسام، م. و هزارجریبی، ا. ۱۳۸۹. ارزیابی تأثیر پیش پردازش پارامترهای ورودی به شبکه عصبی مصنوعی (ANNs) با استفاده از روشهای رگرسیون گام به گام و گاما تست به منظور تخمین سریع تر تبخیر و تعرق روزانه. آب و خاک. ۲۴(۳): ۶۲۴-۶۱۰.

گودرزی، م.، صلاحی، ب. و حسینی، س. ۱۳۹۴. بررسی تغییرات اقلیمی بر تغییرات رواناب سطحی (مطالعه موردی: حوضه آبریز دریاچه ارومیه)، نشریه اکوهیدرولوژی. ۲(۲): ۱۷۵-۱۸۹.

میرعربی، ع.، ناصری، ع.، نخعی، م. و علیجانی، ف. ۱۳۹۷. بررسی عملکرد مدل های داده مینا در شبیه سازی گام‌های مختلف زمانی سطح آب زیرزمینی با استفاده از روش تلفیقی آزمون گاما و الگوریتم ژنتیک. زمین شناسی کاربردی پیشرفته. ۸(۲): ۶۲-۷۲.

Abbas, F., Afzaal, H., Farooque, A.A. and Tang, S. 2020. Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*. 10(7): 1046.

Afrin, S., Khan, A. T., Mahia, M., Ahsan, R., Mishal, M. R., Ahmed, W., and Rahman, R. M. 2018, June. Analysis of soil properties and climatic data to predict crop yields and cluster different agricultural regions of Bangladesh. In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) (pp. 80-85). IEEE.

Ali, M. and Prasad, R. 2019. Significant wave height forecasting via an extreme learning machine model integrated with improved complete ensemble empirical mode decomposition. *Renewable and Sustainable Energy Reviews*. 104: 281-295.

Amirshayeri, A., Behmanesh, J., Rezaverdinejad, V. and Fathollahzadeh Attar, N. 2023. Evapotranspiration estimation using hybrid and intelligent methods. *Soft Computing*. 27(14): 9801-9821.

Arshad, S., Kazmi, J. H., Javed, M. G. and Mohammed, S. 2023. Applicability of machine learning techniques in predicting wheat yield based on remote sensing and climate data in Pakistan, South Asia. *European Journal of Agronomy*. 147: 126837.

Azar, N. A., Kardan, N. and Milan, S. G. 2023. Developing the artificial neural network-evolutionary algorithms hybrid models (ANN-EA)

محدوده، جنس خاک، انواع خصوصیات خاک و همچنین پوشش گیاهی محدوده در این تحقیق به دلیل نبود اطلاعات در نظر گرفته نشده بود، لذا در تحقیقات آتی به عنوان چشم‌انداز این تحقیق می‌تواند در نظر گرفته شود. همچنین تأثیر تغییر اقلیم بر میزان عملکرد گندم دیم می‌تواند تأثیر بسزایی بگذارد، که توصیه می‌گردد جهت تکمیل این تحقیق در تحقیقات آتی در نظر گرفته شوند. همچنین استفاده از شاخص‌های خشکسالی و استفاده از ترکیب مدل‌های یادگیری ماشین با الگوریتم‌های فراابتکاری توصیه می‌گردد در نظر گرفته شود. عدم قطعیت از جمله موارد دیگری است که در این تحقیق در نظر گرفته نشده است لذا در نظر گرفتن آن در تحقیقات آتی توصیه می‌گردد.

تشکر و قدردانی

بدینوسیله از سازمان جهاد کشاورزی استان آذربایجان غربی به دلیل همکاری صمیمانه در ارائه داده‌های مورد نیاز این تحقیق تشکر و قدردانی می‌گردد.

منابع

امیرعشایری، ا.، بهمنش، ج.، رضا وردی نژاد، وحید. و فتح اله زاده، ن. ۱۳۹۹. پیش‌بینی تبخیر-تعرق مرجع روزانه با استفاده از روش ترکیبی هوشمند مصنوعی بر پایه الگوریتم پیش‌پردازش‌کننده تجزیه مد تجزیه. نشریه آبیاری و زهکشی ایران. ۱۴(۴): ۱۳۷۶-۱۳۹۰.

بسالت پور، ع. ا.، حاج عباسی، م. ع. و ایوبی، ش. ۱۳۹۲. استفاده از آزمون گاما برای انتخاب ورودی‌های بهینه در مدل‌سازی مقاومت برشی خاک با استفاده از شبکه‌های عصبی مصنوعی. *مجله پژوهش‌های حفاظت آب و خاک*. ۲۰(۱): ۹۷-۱۱۴.

جعفری، م. م.، لیاقت، ع. م. و محتشمی، س. ۱۴۰۲. مقایسه و ارزیابی الگوریتم‌های LM، BR و GD شبکه عصبی مصنوعی در برآورد عملکرد گندم دیم براساس پارامترهای هواشناسی (مطالعه موردی: استان کرمانشاه). نشریه آبیاری و زهکشی ایران. ۱۷(۳): ۵۴۴-۵۲۹.

جلالی، م.، محمدی، غ. ح.، حسینی صدر، ع. و خوشوقتی، ج. ۱۳۹۶. نقش پراکندگی مکانی و نوسانات زمانی پارامترهای اقلیمی در عملکرد گندم دیم. *جغرافیای طبیعی*. ۹(۳۴): ۱۰۵-۱۲۳.

سیفی، ا.، میر لطیفی، س. م. و ریاحی، ح. ۱۳۸۹. توسعه مدل ترکیبی رگرسیون چندگانه-تحلیل مؤلفه‌ها و عامل‌های اصلی (MLR-PCA) در پیش‌بینی تبخیر و تعرق مرجع (مطالعه

- 995.
- Huang, Y., Yang, L., Liu, S. and Wang, G. 2019. Multi-Step Wind Speed Forecasting Based On Ensemble Empirical Mode Decomposition, Long Short Term Memory Network and Error Correction Strategy. *Energies*. 12(10): 1822.
- Jamnani, M.R., Kayhomayoon, Z., Azar, N.A., Milan, S.G., Marghmaleki, S.N. and Berndtsson, R. 2024. Large discrepancy between future demand and supply of agricultural water in northwestern Iran; evidence from WEAP-MODFLOW-machine learning under the CMIP6 scenario. *Computers and Electronics in Agriculture*. 216:108505.
- Kayhomayoon, Z., Ghordoyee-Milan, S., Jaafari, A., Arya-Azar, N., Melesse, A.M. and Moghaddam, H.K. 2022. How does a combination of numerical modeling, clustering, artificial intelligence, and evolutionary algorithms perform to predict regional groundwater levels?. *Computers and Electronics in Agriculture*. 203:107482.
- Kemp, S. E., Wilson, I. D., and Ware, J. A. 2004. A tutorial on the gamma test. *International Journal of Simulation: Systems, Science and Technology*. 6(1-2): 67-75.
- Khaki, S., Wang, L., and Archontoulis, S. V. 2020. A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*. 10:1750.
- Khouban, L., Ghaiyoomi, A.A., Teshnehlab, M., Ashlaghi, A.T., Abbaspour, M. and Nassiri, P. 2015. COMBINATION OF ARTIFICIAL NEURAL NETWORKS AND GENETIC ALGORITHM-GAMMA TEST METHOD IN PREDICTION OF ROAD TRAFFIC NOISE. *Environmental Engineering & Management Journal (EEMJ)*. 14(4).
- Kim, N., Ha, K. J., Park, N. W., Cho, J., Hong, S. and Lee, Y. W. 2019. A comparison between major artificial intelligence models for crop yield prediction: Case study of the midwestern United States, 2006–2015. *ISPRS International Journal of Geo-Information*. 8(5):240.
- Koncar, N. 1997. Optimisation methodologies for direct inverse neurocontrol (Doctoral dissertation, University of London).
- Kumar, D., Kumar, Y., Gulati, A. and Kukreja, V. 2022, October. Wheat crop yield prediction using machine learning. In 2022 International Conference on Data Analytics for Business and Industry (ICDABI). 433-437. IEEE.
- Meraj, G., Kanga, S., Ambadkar, A., Kumar, P., Singh, S.K., Farooq, M., Johnson, B.A., Rai, A. and Sahu, N. 2022. Assessing the yield of wheat using satellite remote sensing-based machine learning algorithms and simulation modeling. *Remote Sensing*. 14(13):3005.
- Murakami, K., Shimoda, S., Kominami, Y., Nemoto, M. and Inoue, S. 2021. Prediction of municipality-level to predict the daily evaporation from dam reservoirs. *Engineering with Computers*. 39(2): 1375-1393.
- Breiman, L. 2001. Random forests. *Machine Learning*. 45(1): 5-32.
- Dastorani, M., Mirzavand, M., Dastorani, M. T. and Sadatinejad, S. J. 2016. Comparative study among different time series models applied to monthly rainfall forecasting in semi-arid climate condition. *Natural Hazards*. 81(3): 1811-1827.
- Deo, R. C., Downs, N., Parisi, A. V., Adamowski, J. F., and Quilty, J. M. 2017. Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle. *Environmental research*. 155: 141-166.
- Durrant, P. J. 2001. WinGamma: A non-linear data analysis and modelling tool with applications to flood prediction. Unpublished PhD thesis, Department of Computer Science, Cardiff University, Wales, UK.
- Fijani, E., Barzegar, R., Deo, R., Tziritis, E. and Konstantinos, S. 2019. Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters. *Science of the total environment*. 648: 839-853.
- Friedman, J. H. 1991. Multivariate adaptive regression splines. *The annals of statistics*. 19(1): 1-67.
- Gandhi, N., Petkar, O., and Armstrong, L. J. 2016, July. Rice crop yield prediction using artificial neural networks. In 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR). (pp. 105-110). IEEE.
- Ghaemi, A., Rezaie-Balf, M., Adamowski, J., Kisi, O. and Quilty, J. 2019. On the applicability of maximum overlap discrete wavelet transform integrated with MARS and M5 model tree for monthly pan evaporation prediction. *Agricultural and Forest Meteorology*. 278, 107647.
- Grinberg, N. F., Orhobor, O. I. and King, R. D. 2020. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Machine Learning*. 109(2): 251-277.
- Huang, N. E., Wu, M. L., Qu, W., Long, S. R. and Shen, S. S. 2003. Applications of Hilbert–Huang transform to non-stationary financial time series analysis. *Applied stochastic models in business and industry*. 19(3): 245-268.
- Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C. and Liu, H.H. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*. 454(1971): 903-

- Sinwar, D., Dhaka, V. S., Sharma, M. K. and Rani, G. 2020. AI-based yield prediction and smart irrigation. *Internet of Things and Analytics for Agriculture*. 2:155-180.
- Sunil, G. L., Nagaveni, V., and Shruthi, U. 2022, July. A review on prediction of crop yield using machine learning techniques. In *2022 IEEE region 10 symposium (TENSYP)* (pp. 1-5). IEEE.
- Taheri, M., Emadzadeh, M., Gholizadeh, M., Tajrishi, M., Ahmadi, M. and Moradi, M. 2019. Investigating the temporal and spatial variations of water consumption in Urmia Lake River Basin considering the climate and anthropogenic effects on the agriculture in the basin. *Agricultural water management*. 213:782-791.
- Vapnik, V. 1998. *Statistical learning theory*. John Wiley & Sons google schola. 2:831-842.
- Wu, Z. and Huang, N. E. 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in adaptive data analysis*. 1(01):1-41.
- Yao, X. 1999. Evolving artificial neural networks. *Proceedings of the IEEE*. 87(9):1423-1447.
- Yeh, J. R., Shieh, J. S. and Huang, N. E. 2010. Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method. *Advances in adaptive data analysis*. 2(02):135-156.
- winter wheat yield based on meteorological data using machine learning in Hokkaido, apan *PLoS One*. 16 (10):1-19.
- Oshiro, T. M., Perez, P. S. and Baranauskas, J. A. 2012. How many trees in a random forest? In *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8* (pp. 154-168). Springer Berlin Heidelberg.
- Rehman, N. and Mandic, D. P. 2010. Multivariate empirical mode decomposition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 466(2117):1291-1302.
- Remesan, R. and Mathew, J. 2015. *Hydrological Data-Driven Models (A Case Study Approach)*, Springer International Publishing Switzerland, 249P.
- Rezaie-Balf, M., Kisi, O. and Chua, L. H. 2019. Application of ensemble empirical mode decomposition based on machine learning methodologies in forecasting monthly pan evaporation. *Hydrology Research*. 50(2):498-516.
- Rezaie-Balf, M., Naganna, S.R., Kisi, O. and El-Shafie, A. 2019. Enhancing streamflow forecasting using the augmenting ensemble procedure coupled machine learning models: case study of Aswan High Dam. *Hydrological Sciences Journal*. 64(13):1629-1646.
- Rigatti, S. J. 2017. Random forest. *Journal of Insurance Medicine*. 47(1):31-39.

Modeling and Forecasting Rainfed Wheat Yield Based on Meteorological Variables Using Combined Artificial Intelligence Methods

A. Amirashayeri¹, V. Rezaverdinejad², J. Behmanesh^{*3}, F. Asadzadeh⁴, M. Rahimi⁵

Received: Jan.14, 2025

Accepted: Apr.01, 2025

Abstract

The wheat has a vital role in the country food security and its yield estimation can be useful in the regional and macro decisions. To predict the crop yield, artificial intelligence models are known as one of the most suitable tools. Therefore, in the present research a wide range of machine learning models including artificial neural networks (ANN), random forest (RF), support vector regression (SVR) and multivariate adaptive regression spline (MARS) were evaluated in West Azerbaijan where is the most important region in Iran in wheat production. Based on the effective variables on the rainfed wheat yield, including total precipitation during the growing season, number of days prior the first and end rainfall events more than 10mm, relative humidity, evaporation, average sunshine hours and number of frost days were utilized as input variables. The gamma test (GT) method was used in combination with the genetic algorithm (GA) to optimize the number of input variables. Using the gamma test and genetic algorithm, 8 variables were selected as inputs for the models input. Additionally, the models were combined with Ensemble empirical mode decomposition (EEMD) to improve prediction accuracy. The results of this study show The EEMD-MARS model provided the most accurate results, with error evaluation criteria RMSE= 0.112(ton.ha⁻¹), MAE= 0.088(ton.ha⁻¹), NSE= 0.945 and SI= 0.101 for the test stage. Also, in this model, 14 different functions were extracted for estimating rainfed wheat yield. The EEMD-SVR model also had good performance whit comparing other models with error criteria RMSE= 0.132(ton.ha⁻¹), MAE= 0.080(ton.ha⁻¹) and NSE= 0.923 in the test stage.

Keywords: Rainfed products, Ensemble empirical mode decomposition, Gamma test, Machine learning

1 - Ph.D. Candidate of Drainage Irrigation, Department of Water Engineering, University of Urmia, Urmia, Iran

2 - Professor of Water Engineering, Department of Water Engineering, University of Urmia, Urmia, Iran

3 - Professor of Water Engineering, Department of Water Engineering, University of Urmia, Urmia, Iran

4 - Professor of Soil and Water Conservation ,Soil Science Department, University of Urmia, Urmia, Iran

5 - Ph.D. of Water Engineering, Department of Water Engineering, University of Urmia, Urmia, Iran

(* - Corresponding Author Email: j.behmanesh@urmia.ac.ir)