University of Qom
Engineering Management and Soft Computing
JEMSC.QOM
Online ISSN: 2538-2675

# The Diagnosis of Diabetes Using a Hybrid Algorithm Consisting of the Flower Pollination Algorithm and an Ensemble of a Subset of K-NN Classifiers

**Zeinab Hassani**[1]✉ **and Najmeh Samadiani**[2]

1. Correspnding author, Instructor, Faculty of Engineering, Kosar University, Bojnord, Iran. Email: hasani_uni@yahoo.com
2. Instructor, Faculty of Engineering, Kosar University, Bojnord, Iran. Email: najmeh_sam@yahoo.com

**Article Info**

**Article type:**
Research Article

**Article history:**
Received 2022 Jun 6
Received in revised form 2022 Jul 11
Accepted 2022 Jul 13
Published online 2022 Sep 16

**Keywords:**
Diabetes,
Ensemble of a Subset of K-Nearest Neighbor Classifiers,
Flower Pollination Algorithm,
K-Nearest Neighbor Algorithm.

**ABSTRACT**

Diabetes is a disease which, as well as prevention, requires a high level of care, such as monitoring the blood sugar changes. The timely diagnosis of disease plays an important role in its treatment and decreases the damage caused by the disease. Therefore, it is essential to diagnose diabetes. Since hybrid algorithms have a high ability to predict and diagnose various diseases, this article presents an intelligent approach to the diagnosis of this disease, using a hybrid algorithm of flower pollination and K-nearest neighbor ensemble. The accuracy of the proposed method is measured to be 97.78, by using Pima Indians Diabetes (PID) dataset, consisting of 768 samples and 8 features. The results show that the accuracy of this approach has significantly increased compared with the previous studies, and confirms the superiority of the proposed method.

**Cite this article:** Hassani, Z., Samadiani, N. (2022). The Diagnosis of Diabetes Using a Hybrid Algorithm Consisting of the Flower Pollination Algorithm and an Ensemble of a Subset of K-NN Classifiers. *Enginiiring Management and Soft Computing, 8* (1). 37-48. **DOI**: https://doi.org/10.22091/jemsc.2019.1280


DOI: https://doi.org/10.22091/jemsc.2019.1280
**Publisher**: University of Qom

JEMSC.QOM
شاپا الکترونیکی: ۲۶۷۵-۲۵۳۸
مدیریت مهندسی و رایانش نرم
دانشگاه قم

# تشخیص بیماری دیابت با استفاده از الگوریتم ترکیبی گرده‌افشانی گل و الگوریتم گروهی نزدیک‌ترین همسایه

زینب حسنی[۱] و نجمه صمدیانی[۲]

۱. نویسنده مسئول، مربی، گروه مهندسی کامپیوتر، دانشکده علوم پایه و فنی، دانشگاه کوثر، بجنورد، ایران. رایانامه: Hassani@kub.ac.ir
۲. مربی، گروه مهندسی کامپیوتر، دانشکده علوم پایه و فنی، دانشگاه کوثر، بجنورد، ایران. رایانامه: Najmeh_sam@yahoo.com

**اطلاعات مقاله**

**نوع مقاله:** مقاله پژوهشی

**تاریخ دریافت:** ۱۴۰۱/۰۹/۰۸

**تاریخ بازنگری:** ۱۴۰۱/۱۱/۱۲

**تاریخ پذیرش:** ۱۴۰۱/۱۱/۱۴

**تاریخ انتشار:** ۱۴۰۱/۱۲/۲۵

**کلیدواژه‌ها:**
الگوریتم گرده‌افشانی گل،
الگوریتم نزدیک‌ترین همسایه،
الگوریتم گروهی نزدیک‌ترین همسایه،
بیماری دیابت.

**چکیده**

دیابت بیماری است که علاوه بر پیشگیری، نیاز به مراقبت‌های فراوانی از جمله میزان نوسانات سطح قند خون دارد. تشخیص به موقع بیماری نقش بسزایی در درمان ایفا می‌کند و به طور چشمگیری صدمات ناشی از بیماری را کاهش می‌دهد. بنابراین، نیاز به تشخیص بیماری دیابت احساس می‌شود. به دلیل آنکه الگوریتم‌های ترکیبی توانایی بالایی در پیش‌بینی و تشخیص انواع بیماری‌ها دارند، در این مقاله رویکردی هوشمندانه با الگوریتم ترکیبی گرده‌افشانی گل و الگوریتم گروهی نزدیک‌ترین همسایه برای تشخیص این بیماری ارائه شده است. صحت روش پیشنهادی با مجموعه داده PID با ۷۶۸ نمونه و ۸ ویژگی ارزیابی شده و صحت ۹۷/۷۸ درصد به دست آمده است. نتایج نشان می‌دهد که صحت این روش به میزان قابل توجهی نسبت به مطالعات قبلی بهبود یافته است که برتری روش پیشنهادی را تأیید می‌کند.

**استناد:** حسنی، زینب و صمدیانی، نجمه. (۱۴۰۱). «تشخیص بیماری دیابت با استفاده از الگوریتم ترکیبی گرده‌افشانی گل و الگوریتم گروهی نزدیک‌ترین همسایه». *مدیریت مهندسی و رایانش نرم، دوره ۸* (۱). صص: ۴۸-۳۷. https://doi.org/10.22091/jemsc.2019.1280

ناشر: دانشگاه قم

## ۱) مقدمه

دیابت بیماری است که از اختلال متابولیک و افزایش غلظت قند خون ناشی از کمبود انسولین یا حساسیت به انسولین یا هر دو ایجاد می‌شود و می‌تواند منجر به عوارض جدی برای بیمار یا حتی مرگ زودرس آن گردد. با این حال، انجام چندین آزمایش برای تشخیص دیابت ضروری است و تجزیه و تحلیل عوامل بحرانی نقش حیاتی در تشخیص بیماری دارد. بنابراین، امروزه سیستم‌های پزشکی مبتنی بر کامپیوتر (CAD) بسیار مرسوم شده‌اند که از الگوریتم‌های یادگیری ماشین برای طبقه‌بندی و تشخیص بیماری‌ها، درمان به موقع بیماری و کاهش هزینه‌های مورد نیاز استفاده می‌شود (ییلماز، اینان و وستر، ۲۰۱۴). علاوه بر این، این الگوریتم‌ها منجر به تصمیمات دقیق شده و مواردی مانند درد بیمار، زمان‌بر بودن و وابستگی به شرایط محیطی سیستم پیشنهادی را تحت تأثیر در تصمیم‌گیری درمان قرار نمی‌دهند.

## ۲) پیشینهٔ پژوهش

برای تشخیص بیماری دیابت مطالعات گسترده با الگوریتم‌های مختلف یادگیری ماشین انجام شده است. در سال ۲۰۱۳ آنوجا و همکارانش سیستم تشخیص بیماری دیابت با الگوریتم SVM را پیشنهاد داده‌اند و آن را با مجموعه داده PID ارزیابی کردند. صحت سیستم پیشنهادی ۷۸ درصد با تابع کرنل RBF حاصل شده است (آنوجا و چیتر ۲۰۱۳). آیزوریا و همکارانش درخت تصمیم J48 و الگوریتم نایو بیز برای تشخیص دیابت با مجموعه داده PID به کار بردند و نتایج طبقه بندی به ترتیب ۷۴/۸ درصد و ۷۹/۵ درصد برای درخت تصمیم J48 و الگوریتم بیز ساده به دست آمد (آیزوریا، جیالاتا و روناک ۲۰۱۵). هارلن و همکاران، نیز یک سیستم پیشگیری بیماری دیابت مشابه (آیزوریا، جیالاتا و روناک ۲۰۱۵) ارائه دادند که نتایج سیستم پیشنهادی برای درخت تصمیم J48 و الگوریتم بیز ساده به ترتیب ۷۳/۸ درصد و ۷۶/۳ درصد به دست آمده است (هارلین و بامبری ۲۰۱۶). کریشنوینی و همکارانش، روش‌های مختلف از جمله الگوریتم نزدیک‌ترین همسایه KNN، الگوریتم بیز ساده، تحلیل تشخیصی، ماشین بردار با تابع کرنل خطی و تابع کرنل RBF را برای پیش‌بینی بیماری دیابت مطالعه کرده‌اند. بهترین نتیجه در روش تحلیل تشخیصی با صحت ۷۶/۳ درصد حاصل شده است (کریشنوینی و سودا ۲۰۱۷). در سال ۲۰۱۸، به کارگیری الگوریتم ترکیبی الگوریتم‌های فراابتکاری ژنتیک، ازدحام ذرات و گردش کار با کلاس‌بندی شبکه عصبی برای تشخیص بیماری دیابت پیشنهاد شده‌اند که از میان آن‌ها، الگوریتم ازدحام ذرات با شبکه عصبی بالاترین صحت با مقدار ۸۰/۳۴ را نشان داده است (احسان، عثمان، اوغز و خالد ۲۰۱۸).

در این مقاله، روش جدیدی برای تشخیص بیماری دیابت پیشنهاد شده است که قادر است با بالاترین صحت، این بیماری را تشخیص دهد. مبنای الگوریتم‌های کلاس‌بندی، یادگیری ماشین با داده‌های متوازن است و داده‌های پزشکی از جمله داده بیماری دیابت نامتوازن هستند. بنابراین، در این پژوهش از روش بیش نمونه برداری Smote برای متوازن‌سازی داده‌ها استفاده شده است. سپس، عملکرد الگوریتم ترکیبی گرده‌افشانی گل و الگوریتم گروهی نزدیک‌ترین همسایه با داده‌های متوازن شده در مجموعه داده PID بررسی می‌شود.

در ادامه، در بخش دوم روش تحقیق شرح داده می‌شود. نتایج و یافته‌ها بخش سوم را شکل می‌دهند. در بخش چهارم، نتیجه‌گیری پایان‌بخش مقاله خواهد بود.

## ۳) روش‌شناسی پژوهش

در این مطالعه، الگوریتم ترکیبی گرده‌افشانی گل با الگوریتم گروهی نزدیکترین همسایه با مجموعه داده PID[1] مطالعه شده است. ابتدا مجموعه داده PID پیش‌پردازش می‌شود. سپس داده‌های پیش‌پردازش شده را با روش پیشنهادی ارزیابی می‌کنیم. در این بخش، مفاهیم اولیه از جمله الگوریتم گرده‌افشانی گل، الگوریتم گروهی نزدیک‌ترین همسایه و روش پیشنهادی توضیح داده می‌شود.

## ۴) الگوریتم گرده‌افشانی گل

الگوریتم گرده‌افشانی گل (FPA[2]) یا به بیان دیگر، الگوریتم گل، توسط شین یانگ در سال ۲۰۱۲ معرفی شد (لنین ردی ۲۰۱۴) که عملکرد آن بر اساس روند گرده افشاندن گل گیاهان است. گرده‌افشانی پدیده‌ای است که گرده از گلی به گل دیگر انتقال داده می‌شود. برای طراحی الگوریتم FPA، ۴ بخش اصلی مد نظر قرار می‌گیرد.

بخش اول: گرده‌افشانی زیست‌محیطی (biotic) و متقابل (cross) به عنوان روند گرده‌افشانی جهانی مورد توجه قرار گرفته است. گرده‌افشانی جهانی را می‌توان به صورت رابطه (۱) نشان داد.

$$x_i^{t+1} = x_i^t + \gamma L(\lambda)(g_* - x_i^t)$$ رابطه ۱)

$x_i^t$ گرده i یا حل مسئله $x_i$ در تکرار tام، $g_*$ بهترین راه‌حل فعلی در میان تمام راه‌حل‌های موجود در نسل، i حالت فعلی و $\gamma$ یک معیار اندازه برای کنترل اندازه گام‌ها است. $L(\lambda)$ پارامتری مربوط به قدرت گرده‌افشانی است که باعث بهبود فرایند گرده‌افشانی می‌گردد. از آنجا که حشرات ممکن است با گام‌های فاصله مختلف حرکت کنند، از توزیع $L(\lambda)$ برای توزیع پروازها با مقدار بزرگتر از صفر L> 0 استفاده می‌شود.

$$L(\lambda) \sim \frac{\lambda\Gamma(\lambda \sin(\pi\frac{\lambda}{2}))}{\pi} \frac{1}{s^{1+\lambda}} \quad (s \gg s_0 > 0)$$ رابطه ۲)

$\Gamma(\lambda)$ توزیع استاندارد گاما است و توزیع برای مراحل بزرگ s> 0 معتبر است.

بخش دوم: برای گرده‌افشانی محلی، A-biotic و self-pollination استفاده شده است.

بخش سوم: گرده‌افشانی توسط حشرات و پرندگان می‌تواند قابلیت افشانی گل را افزایش دهد که باعث توسعه و ثبات گل‌ها و تولید گونه‌های جدید می‌شود.

بخش دوم و سوم با رابطه (۳) نشان داده شده است.

$$x_i^{t+1} = x_i^t + \mathrm{e}(x_j^t - x_k^t)$$ رابطه ۳)

$x_j^t$ و $x_k^t$ گرده گل‌های متفاوت از گونه‌های گیاهی مشابه است که قابلیت گرده‌افشانی گل را در منطقه محدودی تقلید می‌کنند. از دیدگاه ریاضی، اگر $x_j^t$ و $x_k^t$ از یک گونه یا از همان جمعیت انتخاب شده باشد، رابطه ۳ با $\mathrm{e} \in [۰،۱]$ یک جستجوی محلی در توزیع یکنواخت است.

[1] Pima Indians Diabetes
[2] Flower Pollination Algorithm

بخش چهارم: ارتباط گرده‌افشانی محلی و گرده‌افشانی جهانی را می‌توان با کنترل احتمال $P \in [0, 1]$ با یک گرایش جزئی نسبت به گرده‌افشانی محلی کنترل کرد (عبدالعزیز، علی وعبدالعزیم ۲۰۱۶).

اثربخشی الگوریتم FPA برای حل مسئله به دو دلیل قابل توجیه است: اولین دلیل این است که گرده‌افشانی‌های حشرات می‌توانند در فاصله‌های طولانی حرکت کنند که باعث می‌شود الگوریتم FPA از جستجوی محلی در فضای بسیار بزرگ (کاوش) جلوگیری کند و دلیل دیگر این است که الگوریتم FPA تضمین می‌کند که همواره گونه‌های مشابه از گل‌ها انتخاب می‌شوند که تضمین‌کننده همگرایی سریع و یافتن راه‌حل بهینه (بهره‌برداری) است (هارونا و همکارانش، ۲۰۱۵).

## ۵) الگوریتم نزدیکترین همسایه

الگوریتم نزدیکترین همسایه KNN یکی از الگوریتم‌های طبقه‌بندی یادگیری ماشین و نوعی یادگیری مبتنی بر نمونه است. طبقه‌بندی این روش صرفاً بر اساس تشابه صورت می‌گیرد. فرایند طبقه‌بندی KNN از یک مجموعه داده با تعداد مشخصی ویژگی شروع می‌شود. مجموعه داده‌ها به دو مجموعه آموزش برای ورودی الگوریتم و تست برای ارزیابی صحت الگوریتم تقسیم می‌شوند. برای طبقه‌بندی، داده‌های تست متعلق به کلاسی است که بیشترین آراء را در بین k نزدیک‌ترین همسایگان آن داشته باشد. برای بدست آوردن نزدیک‌ترین همسایگان یک نمونه، فاصله‌های مختلفی همچون فاصله اقلیدسی و منهتن طبق رابطه (۴) و (۵) استفاده می‌شود (آلکا دمراندر ۲۰۱۶) و (بوانساری، برنتا ۲۰۱۵).

رابطه ۴) $(x, t) = \sqrt{\sum_{1 \le i \le n} (x_i - t_i)^2}$

رابطه ۵) $d(x, t) = \sum_{1 \le i \le n} |x_i - t_i|$

## ۶) الگوریتم گروهی نزدیک‌ترین همسایه

الگوریتم گروهی، ترکیب چندین الگوریتم کلاس‌بندی است که به عنوان بهبود عملکرد کلاس‌بندی‌های با یادگیری ضعیف ارائه شده است (اسما و همکارانش، ۲۰۱۶). طبقه‌بندی بهبودیافته را می‌توان با استفاده از طبقه‌بندی‌های متنوع استفاده کرد. Bagging، Boosting و Rotation Forest سه روش شناخته شده از الگوریتم‌های گروهی هستند.

Bootstrap aggregation (bagging) یکی از ساده‌ترین تکنیک‌های گروهی است. خروجی‌های حاصل از طبقه‌بندی‌هایی که روی مجموعه‌های آموزش bootstrap به طور تصادفی ایجاد شده‌اند را با هم ترکیب می‌کند. از الگوریتم‌های کلاس‌بندی متنوعی مانند درخت تصمیم به عنوان طبقه‌بندهای پایه این الگوریتم استفاده می‌شود. الگوریتم KNN یکی دیگر از این الگوریتم‌های کلاس‌بندی است. در الگوریتم گروهی نزدیک‌ترین همسایه، هدف بهبود دقت طبقه‌بند KNN با استفاده از الگوریتم گروهی است. الگوریتم گروهی KNN، زیرمجموعه‌ای از کلاس‌بندی KNN است که در آن مجموعه داده‌ها به دو مجموعه آموزش و تست تقسیم بندی می‌شود. همچنین، در هر الگوریتم کلاس‌بندی KNN زیرمجموعه‌ای از ویژگی‌ها به صورت تصادفی انتخاب می‌شوند که تکنیک bootstrap با m کلاس‌بند لحاظ می‌شود. فرایند الگوریتم گروهی با زیرمجموعه‌ای از کلاس‌بندی KNN در ادامه بیان شده است. با فرض این که n تعداد نمونه در مجموعه داده L با d ویژگی در دو کلاس است.

۱) مجموعه داده $L^*$ با l ویژگی تصادفی از مجموعه داده L با d ویژگی، بدون جایگزینی انتخاب می‌شود (l<d).

۲) تعدادی نمونه تصادفی از مجموعه داده $L^*$ انتخاب می‌شود.

۳) طبقه‌بند KNN برای نمونه تصادفی انتخابی محاسبه می‌شود.

۴) دقت طبقه‌بند KNN بر اساس تکنیک OOB (در تکنیک OOB یک سوم از داده‌ها برای تست و مابقی برای آموزش مدل طبقه‌بندی استفاده می‌شود) محاسبه می‌شود.

۵) مراحل ۱ تا ۴ به اندازه m تکرار شده و مدل طبقه‌بندی بر اساس دقت مدل مرتب‌سازی می‌شود.

۶) h مدل طبقه‌بندی با دقت‌های بالا انتخاب می‌شوند.

انتخاب مدل طبقه‌بندی به این صورت است که در الگوریتم گروهی دو مدل کلاس‌بندی با دقت بالا لحاظ شده و ارزیابی بر اساس داده‌های validation محاسبه می‌شود. سپس، سومین مدل کلاس‌بندی بعدی اضافه شده و الگوریتم گروهی ارزیابی می‌شود تا h کلاس‌بند انتخاب شود.

در مرحله اول، نرخ خطای کلاس‌بندی به عنوان ارزیاب عملکرد مدل کلاس‌بند بررسی می‌شود و مدل کلاس‌بندی‌ها با نرخ خطای کمتر انتخاب می‌شوند. در مرحله بعدی الگوریتم گروهی، از Brier score برای ارزیابی عملکرد مدل کلاس‌بندی استفاده می‌شود. در این حالت علاوه بر دقت پیش‌بینی مدل کلاس‌بندی، کیفیت مدل نیز بررسی می‌شود. مقدار عملکرد الگوریتم گروهی با r-1 مدل BS(r-1) و r مدل BS(r) است. الگوریتم گروهی r مدل در صورتی که رابطه $BS(r-1) > BS(r)$ برقرار باشد، انتخاب می‌شود. مقدار Brier score برای داده‌های تست با دو تا کلاس {0,1} برای پیش‌بینی کلاس ۱ از رابطه (۶) به دست می‌آید.

$$S = E(y_i - p(y_i = 1))^2$$ رابطه ۶)

براورد مقدار عملکرد BS بالا برابر است با:

$$\widehat{BS} = \sum_{i=1}^{n_t}(y_i - \hat{p}(y_i|x))^2/n_t$$ رابطه ۷)

$n_t$ تعداد کل داده‌های تست است. مدلی کاراتر است که BS کمتری را داشته باشد. پس در الگوریتم گروهی BS کمتر انتخاب می‌شوند. شبه کد الگوریتم گروهی KNN در الگوریتم ۱ بیان شده است (اسما و همکارانش، ۲۰۱۶).

**الگوریتم ۱: شبه کد الگوریتم گروهی نزدیکترین همسایه**

**Generate m models**
1: Randomly split the data into two parts, learning part and validation part.
2: **For** i=1 →m **do**
3: Select a random sample of l features out of total of features
4: Select a random sample of size n with replacement from the learning part for producing $B_i$, where $B_i$ is the bootstrap sample used for constructing the ith model based on l features.
5: Save the instances left out from the ith bootstrap sample as OOB(i)
6: Call build KNN(K) to construct model $C_i$
7: Find accuracy of Ci using OOB(i) and store it as Acc(i)
8: **end for**
9: **Select the best models on the basis of individual performance**
10: **for** j=1→m **do**
11: **If** Acc(j)>$Q_2$ **then**
12: Slect $C_j$, where $Q_2$ is the second quartile of the accuracies of all h models
13: **else**
14: Drop Cj
15: **end if**
16: **end for**
17: Build KNN(k)
18: Construct Knn using a bootstrap sample with l selected features
19: **Return** (Selected models)
20: **Fusing the best models based on collective performance**
21: Arrsnge selected models, say h, in decreasing order with respect to their accuracy
22: Initialize q=1 and take the best model with the highest accuracy from the above selected models as the starting ensemble.
23: f**or** q=2→h **do**
24: **if** $BS^{(q)} < BS^{(q-1)}$ **then**
25: Select the qth knn model, where BS(q) is the brier score of the ensemble having the qth model and BS(q-1) is the brier score of the ensemble havingthe qth model and BS(q-1) is the Brier score of the ensemble not having the gth model afther the validation data
26: **else**
27: Do not selecte the $q^{th}$ model for final ensemble
28: **end if**
29: **end for**

## ۷) متوازن سازی Smote

یکی از روش‌های متوازن‌سازی در داده‌های نامتوازن، روش بیش نمونه‌برداری است. در روش بیش نمونه‌برداری برای به تعادل رساندن توزیع کلاس از روش جایگزینی نمونه‌های کلاس اقلیت استفاده می‌شود تا مجموعه داده متوارن شوند.

متوازن‌سازی Smote[3] یکی از تکنیک‌های روش بیش نمونه‌برداری است (فرانسیسکو، جوزه، جورج و جان ۲۰۱۸). Smote در سال ۲۰۰۲ به عنوان یک رویکرد هوشمند بیش نمونه‌برداری معرفی شده است. در روش بیش نمونه‌برداری، داده‌های کلاس اقلیت به صورت تصادفی جایگزین می‌شوند. در مقایسه با روش‌های بیش نمونه‌برداری تصادفی، روش Smote به طور مؤثر در الگوریتم‌های کلاس‌بندی مانع از بیش‌برازش می‌شود. ویژگی‌های نمونه‌های کلاس اقلیت مشابه است. بر این اساس، با استفاده از تکنیک الگوریتم نزدیک‌ترین همسایه به صورت تصادفی نمونه‌های جدید را به کلاس اقلیت اضافه می‌کند. علاوه بر آن، ویژگی‌ها را با مقدار تصادفی بین مقادیر ۰ و ۱ ضرب می‌کند که باعث افزایش تنوع نمونه‌های جدید می‌شود و امکان بهره‌برداری بهتر از فضای تصمیم‌گیری داده را می‌دهد (مورالی و جورج، ۲۰۰۷؛ زهنگ، ۲۰۱۵).

[3] Synthetic Minority Oversampling Technique

## ۸) مجموعه داده

مجموعه داده‌های PID از مجموعه داده‌های UCI انتخاب شده است. داده‌ها مربوط به ۷۶۸ زن حداقل ۲۱ ساله هندی می‌باشد که ۸ ویژگی برای هر رکورد استخراج شده است. در این میان داده‌های مربوط به ۵۰۰ نفر سالم و ۲۶۸ نفر مبتلا به بیماری دیابت وجود دارد. ۸ ویژگی بر اساس تعریف سازمان جهانی بهداشت ثبت شده است که جدول ۱ ویژگی‌های مجموعه داده PID را نمایش می‌دهد (پایگاه داده یادگیری ماشین UCI).

**جدول ۱. ویژگی‌های مجموعه داده PID**

| ردیف | | ویژگی‌ها |
|---|---|---|
| ۱ | تعداد هفته‌های بارداری | pregnancy Count |
| ۲ | میزان تجمع گلوکز در پلاسمای خون | Glucose concentration in plasma |
| ۳ | فشار خون | Blood pressure (diastolic, mm Hg) |
| ۴ | میزان پوست triceps برای تخمین چربی بدن | Thickness of triceps skin fold (mm) |
| ۵ | انسولین سرم ساعتی | Hour serum insulin (µ U/ml) |
| ۶ | شاخص جرمی بدن | Body mass index |
| ۷ | بررسی سابقه خانوادگی دیابت | pedigree function of diabetes |
| ۸ | سن | Years of age |

## ۹) روش پیشنهادی

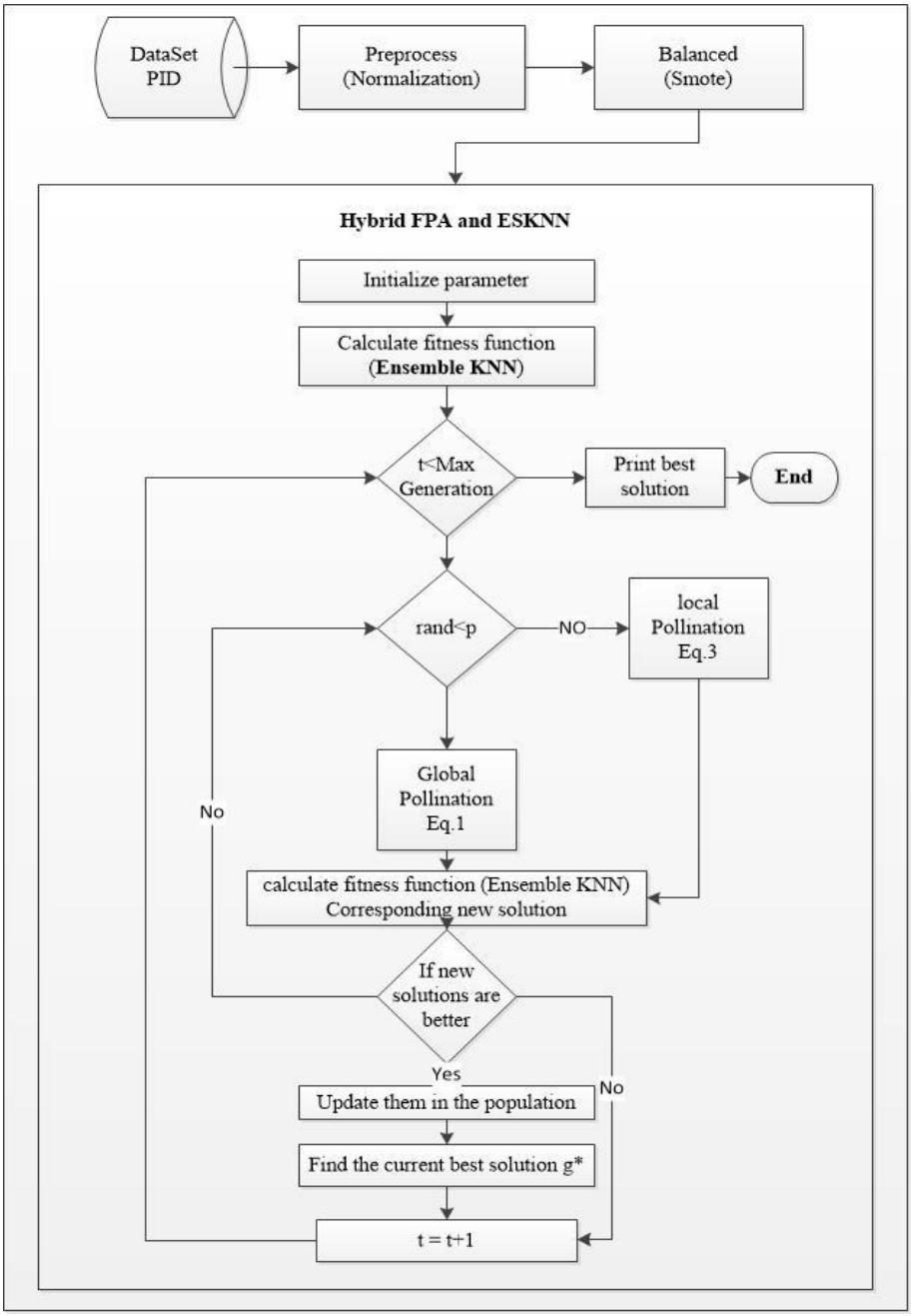
در این بخش روش پیشنهادی برای پیش‌بینی بیماری دیابت شرح داده می‌شود. شکل ۱ الگوریتم پیشنهادی را نمایش می‌دهد.

مرحله اول پیش پردازش داده است که تولید نتایج کارا، به داده‌های مناسب درسیستم پیشنهادی وابسته است. در این مرحله برای یکپارچه‌سازی داده‌ها، آن‌ها را نرمال‌سازی می‌کنیم. با استفاده از رابطه (۱۲) داده‌ها در بازه [0,1] نرمال‌سازی می‌شوند.

$$X = (x - x_{MIN})(x_{MAX} - x_{MIN})$$ رابطه ۱۲)

$(X_{max})$, $(X_{min})$ به ترتیب کمترین و بیشترین مقدار در هر متغیر یا ویژگی هستند.

در مرحله دوم برای متوازن‌سازی داده‌ها از تکنیک Smote استفاده شده است.




**شکل ۱. الگوریتم گرده‌افشانی گل با الگوریتم گروهی نزدیک‌ترین همسایه**

از طرفی در الگوریتم‌های یادگیری ماشین، فرض بر این است که توزیع کلاس‌ها متوازن باشد. از این‌رو در صورت استفاده از این الگوریتم‌ها در طبقه‌بندی داده‌های نامتوازن مدل بدست آمده به سمت نمونه‌های آموزشی کلاس بزرگ‌تر متمایل می‌شود که سبب کاهش دقت مدل حاصل در پیش‌بینی کلاس اقلیت می‌شود و چالش اصلی شناسایی صحیح نمونه های کلاس اقلیت است. با توجه به این که داده‌های معرفی شده نامتوازن هستند، برای بالا بردن دقت روش پیشنهادی باید داده‌ها را متوازن کنیم.

در مرحله سوم از الگوریتم گرده‌افشانی گل با الگوریتم گروهی نزدیک‌ترین همسایه برای پیش‌بینی بیماری دیابت استفاده شده است. در ادامه، شبه کد الگوریتم ترکیبی گرده‌افشانی گل با الگوریتم گروهی نزدیک‌ترین همسایه بیان می‌شود. در جدول ۲ مقادیر اولیه پارامترهای روش پیشنهادی نمایش داده شده است.

**جدول ۲. مقادیر اولیه پارمترهای روش پیشنهادی**

| پارامتر | مقدار |
|---|---|
| احتمال کنترل(probabibility switch) | ۰/۶ |
| حداکثر تعداد نسل‌ها | ۲۰۰، ۵۰۰ |
| جمعیت اولیه | ۱۰، ۲۰ و ۲۵ |

**الگوریتم ۲. الگوریتم گرده‌افشانی گل با الگوریتم گروهی نزدیک‌ترین همسایه**

```
1:  Initialize parameter (population n, control probability P, iterations t, define x = (x1, x2,…, xd)
2: calculate fitness function (Ensemble KNN)
3:  Find the best solution g*
4: while (t<Max Generation)
5: for i = 1:n
6:   if rand<p,
     Draw a (d-dimensional) step vector L which obeys a Levy Distribution Global pollination through
8: else
       Draw ∈ from a uniform distribution in [0,1]
       Do local pollination through
     // end if
9:  calculate fitness function (Ensemble KNN) (Evaluate new solutions)
10:  If new solutions are better,
        Update them in the population
    // end for
11: Find the current best solution g*
12: t = t+1
   //end while
Output: best solution has been found
```

## ۱۰) یافته‌های پژوهش

در این مقاله، الگوریتم گرده‌افشانی گل با الگوریتم گروهی KNN با مجموعه داده PID مورد مطالعه قرار گرفته است. عملکرد مدل پیشنهادی با معیار ارزیابی دقت، صحت، حساسیت و ویژگی بررسی می‌شود. در چندین آزمایش مختلف، با تعداد تکرار و جمعیت متفاوت از داده‌ها عملکرد روش پیشنهادی را ارزیابی می‌کنیم. نتایج نشان می‌دهد که روش پیشنهادی با جمعیت اولیه ۲۰ نمونه و تعداد تکرار ۵۰۰، می‌تواند با بالاترین صحت برابر با ۹۷/۷۸ درصد، دقت ۹۳/۱۰ درصد، ویژگی ۹۶/۹۵ درصد و حساسیت ۹۰ درصد بیماری دیابت را تشخیص دهد. میانگین صحت پردازش‌های انجام شده برابر با ۹۵/۶۹ درصد است که برتری روش پیشنهادی نسبت به مطالعات پیشین را نشان می‌دهد. نتایج پردازش داده‌ها روی مدل پیشنهادی در جدول ۳ نمایش داده شده است.

**جدول ۳. نتایج الگوریتم ترکیبی FPA + Ensemble KNN با مجموعه داده PID**

| جمعیت | تعداد تکرار | ویژگی | حساسیت | دقت | صحت |
|---|---|---|---|---|---|
| ۱۰ | ۲۰۰ | ۹۴/۵۱ | ۷۶/۶۷ | ۹۲ | ۹۴/۹۶ |
| | ۵۰۰ | ۹۶/۳۸ | ۹۶/۱۵ | ۸۳/۳۳ | ۹۶/۳۴ |

| جمعیت | تعداد تکرار | ویژگی | حساسیت | دقت | صحت |
|---|---|---|---|---|---|
| ۲۰ | ۲۰۰ | ۹۶/۳۴ | ۸۳/۳۳ | ۹۶/۱۵ | ۹۶/۳۸ |
| | ۵۰۰ | ۹۶/۹۵ | ۹۰ | ۹۳/۱۰ | ۹۷/۷۸ |
| ۲۵ | ۲۰۰ | ۹۴/۵۱ | ۷۳/۳۳ | ۹۵/۶۵ | ۹۴/۳۳ |
| | ۵۰۰ | ۹۵/۱۲ | ۷۳/۳۳ | ۱۰۰ | ۹۴/۳۷ |

برای نشان دادن برتری روش پیشنهادی، آن را با سایر پژوهش‌های موجود در این حوزه مقایسه کرده‌ایم. جدول ۴ نتایج این مقایسه را نشان می‌دهد. بالاتر بودن دقت روش پیشنهادی، به کارگیری یکی از جدیدترین الگوریتم‌های تکاملی و پیاده‌سازی آسان‌تر به خصوص در مقایسه با شبکه عصبی، از نقاط مثبت مدل پیشنهادی است.

**جدول ٤. مقایسه روش پیشنهادی با با مجموعه داده PID با مطالعات پیشین**

| مراجع | روش کار | صحت |
|---|---|---|
| (آنوجا و چیتر، ۲۰۱۳) | SVM | ۷۸ |
| (آیزوریا، جیالاتا و روناک، ۲۰۱۵) | درخت تصمیم J48 | ۷۴/۸ |
| | Naïve Bayes | ۷۹/۵ |
| (هارلین و بامبری، ۲۰۱۶) | Naïve Bayes و درخت تصمیمJ48 | ۷۶/۳ |
| (کریشنوینی و سودا، ۲۰۱۷) | Discriminant ،Naïve Bayes ،KNN SVM ،analysis | ۷۶/۳ |
| (احسان، عثمان، اوغز و خالد، ۲۰۱۸) | PSO+ANN | ۸۰/۳۴ |
| روش پیشنهادی | FPA + Ensemble KNN | ۹۷/۷۸ |

## ۱۱) نتیجه‌گیری و پیشنهادها

امروزه روش‌های هوش مصنوعی و یادگیری ماشین در علم پزشکی به خصوص در تشخیص بیماری‌ها به کار گرفته می‌شوند. از آنجا که بیماری دیابت بیماری مزمنی است و تشخیص به موقع بیماری در درمان بیماری نقش بسزایی دارد، برآن شدیم تا سیستمی مبتنی بر الگوریتم‌های یادگیری ماشین برای تشخیص و شناسایی این بیماری پیشنهاد دهیم. در این مقاله با ارائه الگوریتم ترکیبی گرده‌افشانی گل و الگوریتم گروهی نزدیک‌ترین همسایه توانستیم بیماری دیابت را با دقت ۹۳ درصد و صحت ۹۷/۷۸ درصد تشخیص دهیم. نتایج حاصل، بالاتر بودن صحت روش پیشنهادی در تشخیص بیماری دیابت نسبت به مطالعات پیشین را نشان می‌دهد.

## منابع


Abdelaziz A.Y., Ali E.S., Abd Elazim S.M., (2016), Flower pollination algorithm to solve combined economic and emission dispatch problems, Engineering Science and Technology, *an International Journal*, 19: 980–990 DOI:10.1007/s10287-009-0113-8

Aiswarya I., S. Jeyalatha and Ronak S., (2015), Diagnosis of Diabetes Using Classification Mining Techniques", *International Journal of Data Mining & Knowledge Management Process (IJDKP)*,5(1): 1-14 DOI:10.1007/s10287-009-0113-8

Alka L., Dharmender K., (2016), Survey on KNN and Its Variants, IJARCCE *International Journal of Advanced Research in Computer and Communication Engineering*, 5(5): 430-435. DOI:10.1109/CEC.2005.1554852

Anuja V. and Chitra R., (2013), Classification Of Diabetes Disease Using Support Vector Machine", *International Journal of Engineering Research and Applications* (IJERA), 3(2): 1797-1801 DOI:10.1016/j.ejor.2006.12.024

Asma G., Aris P., Zardad K., Osama M., Miftahuddin M., Werner A., Berthold L., (2016), Ensemble of a subset of kNN classifiers, *Mathematics Subject Classification*, 1-14. Doi: 10.22091/jemsc.2019.1294

Bhuvaneswari P., Brintha A., (2015), Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm. *Procedia Materials Science*. 10: 433 – 440. DOI:10.1016/j.cor.2005.06.017

Francisco J.C., Jose J.V., Jorge C., Juan R.R., (2018), Oversampling imbalanced data in the string space, Pattern Recognition Letters,103: 32-38. DOI:10.1016/j.eswa.2011.03.060

G. Krishnaveni, T. Sudha, (2017) A Novel Technique to Predict Diabetic Disease Using Data Mining Classification Techniques" in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT- 2017), 3(1): 5-11 Doi: 10.22091/jemsc.2019.1294

Harleen and Bhambri P, (2016), A Prediction Technique in Data Mining for Diabetes Mellitus," *Journal of Management Sciences and Technology*, 4(1): 1-12. Doi: 10.22091/jemsc.2019.1294

Haruna C., Liyana M.S., Sanah A.M., Adamu I. A., et al, (2015), A Review of the Applications of Bio-Inspired Flower Pollination Algorithm, *The 2015 International Conference on Soft Computing and Software Engineer*ing (SCSE 2015), 62: 435-441 Doi: 10.22091/jemsc.2019.1294

Ihsan S., Osman N., Oguz B. and Khalid S., (2018), *Impact of Metaheuristic Iteration on Artificial Neural Network Structure in Medical Data*, Processes, 6, 57. DOI:10.1109/TEVC.2008.925798

Lenin K. and Reddy B. R., (2014), Hybrid Eagle Strategy Flower Pollination Algorithm for Solving Optimal Reactive Power Dispatch Problem, *International Journal of Electrical Energy*, 2(3) DOI:10.1016/j.sbspro.2013.03.036

Lichman, M. UCI Machine Learning Repository; University of California, School of Information and Computer Science: Irvine, CA, USA, http://www.ics.uci.edu/ ~ mlearn/MLRepository.html DOI:10.1016/j.amc.2003.10.057

Murali V., and George S. (2007). An overview of internet addiction. *Advances in Psychiatric Treatment*, 13: 24-30. Doi: 10.1016/j.proeng.2013.04.103

Yilmaz N., Inan O., Uzer M.S, (2014), A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases," *J Med Syst*, 38(5): 38-48. DOI:10.1007/BF02282055

Zheng. Z. (2015). Oversampling method for imbalanced classification computing and Informatics, 34: 1017–1037. Doi: 10.1002/nav.3800030110