Encyclopedia of Digital Transformation
Vol. 2; No. 4; PP. 98-120
winter 2021
DOI: 10.22034/dtj.2023.386891.1084

MehrAlborz University

Presenting a Diabetes Diagnosis Model Based on Recurrent Deep Neural Networks and Oversampling algorithm

Farzaneh Aghamahmoudian[1], Naser Nematbakhsh[2], Mohsen Akhavan Tabib[3]

Abstract

Diabetes is a group of metabolic disorders that are the result of untreated high blood glucose. Early diagnosis and continued control of this disease can reduce its effects. Therefore, providing a method for timely diagnosis of this disease is of great importance. Until now, researchers have made many efforts to provide machine learning methods to diagnose diabetes. But most of these models are either based on simple machine learning methods such as support vector machine or based on the assumption that the available diabetes data are balanced. Both cases are factors of their complete failure. Therefore, considering the existing challenges as well as the importance of timely diagnosis of diabetes, in this research, a diabetes diagnosis model based on deep recurrent neural networks and SMOTE oversampling algorithm is presented. In this model, to improve the diagnosis of diabetes, several pre-processing steps including quantification of missing values, removal of outlier data and then oversampling have been performed. Three deep recurrent neural networks with three recurrent hidden units including LSTM, GRU and BiLSTM have been used to diagnose diabetes. The results of the model presented on the Pima database repository indicate that the average accuracy in 10 different runs in LSTM, GRU and BiLSTM is 91.21%, 89.61% and 90.99%, respectively. The recurrent network with GRU unit has achieved the highest accuracy of 93.74% on average in 10 different executions. The results show that deep neural networks have a much more successful performance in diabetes diagnosis compared to traditional machine learning methods

Keywords: *Diabetes Diagnosis, Recurrent Neural Network, Oversampling, Outlier Data, Machine Learning.*

1.MSc.Student in computer software, Faculty of Technical Engineering, Shahid Ashrafi Esfahani University, Isfahan, Iran

2.Assistant Prof,Computer Department, Faculty of Technical Engineering, Shahid Ashrafi Esfahani University, Isfahan, Iran

3. Ph.D, Department of Health and Medicine, Faculty of Medicine, Islamic Azad University, Najaf Abad, Iran

---

Submitted: 2023-02-25
Accepted: 2023-10-10
Corresponding Author:Farzaneh Aghamahmoudian
Email: farzaneh.am2019@gmail.com

دانشنامهٔ تحول دیجیتال

دانشگاه مهرالبرز
دورهٔ ۲، شمارهٔ ۴
زمستان ۱۴۰۰
صص. ۹۸- ۱۲۰

Encyclopedia of Digital Transformation
DOI: 10.22034/dtj.2023.386891.1084

ارائه یک مدل تشخیص دیابت مبتنی بر شبکه‌های عصبی عمیق بازگشتی و الگوریتم بیش نمونه‌گیری

فرزانه آقامحمودیان[۱]، ناصر نعمت‌بخش[۲]، محسن اخوان طبیب[۳]

چکیده

دیابت، به گروهی از اختلالات متابولیکی گفته می‌شود که نتیجه عدم کنترل قند خون است. تشخیص به موقع و در ادامه کنترل این بیماری به خوبی باعث کاهش اثرات ناشی از آن مثل رتینوپاتی دیابتی، گرفتگی قلبی و عروقی سکته‌های مغزی و غیره می‌شود. محققان تا به امروز تلاش‌های بسیاری در این زمینه کرده‌اند؛ اما اغلب این مدل‌ها یا مبتنی بر روش‌های یادگیری ماشین سنتی هستند و یا بر این فرض استوارند که داده‌های دیابت متوازن هستند. از این رو، در این پژوهش یک مدل تشخیص بیماری دیابت، مبتنی بر شبکه‌های عصبی بازگشتی عمیق و الگوریتم بیش نمونه‌گیری SMOTE ارائه شده است. در این مدل چندین مرحله پیش‌پردازش شامل مقداردهی به مقادیر از دست رفته، حذف داده‌های پرت و سپس بیش نمونه‌گیری انجام شده است. از سه شبکه عصبی عمیق بازگشتی با سه واحد پنهان بازگشتی شامل LSTM, GRU و BiLSTM برای تشخیص استفاده شده است. نتایج مدل ارائه شده بر روی پایگاه داده Pima حاکی از آن است که میانگین صحت در ۱۰ اجرای مختلف در LSTM و GRU و BiLSTM به ترتیب ۹۱/۲۱٪، ۸۹/۶۱٪ و ۹۰/۹۹٪ است. نتایج مدل بازگشتی به ما نشان می‌دهد، شبکه‌های عصبی عمیق در مقایسه با روش‌های یادگیری ماشین سنتی، عملکرد بسیار موفق‌تری دارند.

کلمات کلیدی: *تشخیص دیابت، شبکه عصبی بازگشتی، بیش نمونه‌گیری، داده‌های پرت، یادگیری ماشین*

---

۱. دانشجوی کارشناسی ارشد نرم‌افزار کامپیوتر، دانشکده فنی مهندسی، دانشگاه شهید اشرفی اصفهانی، اصفهان، ایران
۲. استادیار گروه کامپیوتر، دانشکده فنی مهندسی، دانشگاه شهید اشرفی اصفهانی، اصفهان، ایران
۳. دکترای حرفه‌ای پزشکی، گروه بهداشت و درمان، دانشکده پزشکی، دانشگاه آزاد اسلامی واحد نجف‌آباد، نجف‌آباد، ایران

تاریخ دریافت مقاله: ۱۴۰۱/۱۲/۰۶
تاریخ پذیرش نهایی مقاله: ۱۴۰۲/۰۷/۱۸
نویسندهٔ مسئول مقاله: فرزانه آقامحمودیان
Email: farzaneh.am2019@gmail.com

مقدمه

منظور از دیابت، گروهی از اختلالات متابولیکی است که به دلیل عدم کنترل قند خون بروز می‌کند. دیابت سالانه منجر به مرگ تقریباً ۴ میلیون نفر می‌شود (سازمان[۱]، ۲۰۱۹). ابتلا به این بیماری در دهه اخیر به شدت افزایش یافته است که از عوامل اصلی آن کم‌تحرکی مردم به دلیل زندگی مدرن و تغییر در عادات غذایی است. آمارهای ارائه شده توسط سازمان بین‌المللی دیابت[۲] نشان می‌دهد که در حال حاضر بالغ بر ۱ میلیون نوجوان در سراسر جهان مبتلا به دیابت هستند (سازمان، ۲۰۱۹). سه بیماری بسیار خطرناک رتینوپاتی، نفروپاتی و نوروپاتی از جمله بیماری‌های ناشی از دیابت است که به ترتیب بر عروق چشمی، کلیه و اعصاب محیطی مؤثر بوده و منجر به از کار افتادن اندام‌های وابسته می‌شود. علاوه بر این افراد دیابتی در معرض خطر ابتلا به بیماری‌های دیگری مثل بیماری‌های قلبی، شریانی و عروق مغزی، چاقی، آب مروارید و بیماری کبد چرب هستند. طبق دسته‌بندی سازمان بهداشت جهانی[۳]، دو نوع عمده دیابت نوع ۱ و دیابت نوع ۲ در جهان مطرح است. تفاوت این دو نوع دیابت با توجه به فاکتورهایی مثل سن ابتلا، میزان عملکرد سلول بتا، میزان مقاومت به انسولین و نیاز به درمان انسولین برای بقا مشخص می‌شود (سازمان، ۲۰۱۹). انتخاب روش کنترل نامناسب و همچنین تشخیص دیر هنگام این بیماری می‌تواند عوارض مرگباری برای افراد به همراه داشته باشد؛ بنابراین، ارائه روشی مبتنی بر هوش مصنوعی به عنوان دستیار پزشکان برای تشخیص به موقع و انتخاب روش کنترلی مناسب، امری ضروری به نظر می‌رسد. یادگیری ماشین و یادگیری عمیق، ابزاری است که می‌تواند در این زمینه مفید واقع شود. استفاده از فناوری‌های یادگیری ماشین و روش‌های طبقه‌بندی می‌تواند در تشخیص به موقع این بیماری با بررسی سوابق بیمار بسیار مؤثر باشد؛ به نحوی که عوارض ناشی از این بیماری کاهش پیدا کند و کنترل به موقع آن منجر به بهبود کیفیت زندگی فرد شود. روش‌های متعددی تاکنون برای تشخیص بیماری دیابت مبتنی بر فناوری‌های مختلف یادگیری ماشین ارائه شده‌اند. (چاجوری و گوپتا[۴]، ۲۰۱۹)، (فیارنی و همکاران[۵]، ۲۰۱۹)، (گانش و سرپریا[۶]، ۲۰۱۹)، (قربانی و گوسی[۷]، ۲۰۱۹)؛ اما اغلب این روش‌ها مبتنی بر روش‌های یادگیری ماشین سنتی هستند و یا بر این فرض استوارند که داده‌های دیابت متوازن هستند؛ اما داده‌های پزشکی حقیقی متوازن نیستند. در داده‌های پزشکی غالباً یک کلاس اکثریت و کلاس دیگر اقلیت است که عدم توجه به این مورد باعث کاهش کارایی و دقت روش‌های یادگیری ماشین می‌شود که در حوزه پزشکی که بحث حیات بیماران مطرح است، اصلاً مسئله کوچکی نیست (کروچک[۸] ، ۲۰۱۶)، (ژای و همکاران[۹]، ۲۰۱۷). با وجود تلاش‌های متعدد در یادگیری ماشین، طبقه‌بندی داده‌های نامتوازن همچنان از جمله چالش‌های مطرح در این زمینه است. الگوریتم‌های یادگیری ماشین بر این فرض استوار هستند که داده‌های دریافتی دارای کلاس‌های متوازن بوده و دارای اهمیت برابر هستند؛ اما مجموعه داده‌های متوازن در دنیای حقیقی بسیار نادر هستند و داده‌های کلاس اقلیت که دارای نمونه‌های کمتری هستند، بیشترین نرخ طبقه‌بندی نادرست را دارند که این مسئله در حوزه پزشکی بسیار مشکل‌ساز است. به عنوان مثال، در کشور انگلستان در حال حاضر در حدود ۴٪ از مردم مبتلا به دیابت هستند و ۹۶٪ نیز سالم هستند (دیابت[۱۰]، ۲۰۱۵). حال اگر فرض شود از یک روش یادگیری ماشین برای طبقه‌بندی و تشخیص دیابت در این داده‌ها استفاده شود؛ تمام داده‌های کلاس اکثریت (کلاس سالم) را به طور صحیح طبقه‌بندی می‌کند و تمام داده‌های کلاس اقلیت (مبتلا به دیابت) را به اشتباه به کلاس اکثریت تخصیص می‌دهد و اشتباه طبقه‌بندی می‌کند. طبقه‌بندی نادرست در حوزه پزشکی عواقب جبران‌ناپذیری را به دنبال خواهد داشت. اهمیت

1. Organization
2. International Diabetes Federation (IDF)
3. World Health Organization
4. Choudhury & Gupta
5. Fiarni et al.
6. Ganesh & Sripriya
7. Ghorbani & Ghousi
8. Krawczyk
9. Zhai et al.
10. Diabetes.

این موضوع باعث شده است که در سال‌های اخیر، پژوهش‌هایی در زمینه داده‌های نامتوازن، توجه بیشتری را به خود جلب کند (فرناندزو همکاران[1]، ۲۰۱۸). با توجه به اهمیت این موضوع، پژوهش‌های مختلفی برای کنترل عدم توازن داده‌های پزشکی ارائه شده‌اند که مبتنی بر فناوری‌های بازنمونه‌گیری[2] هستند (پرتیوی و همکاران[3]،۲۰۲۰)، (حایرانی و همکاران[4]، ۲۰۲۰)، (سرجیث و همکاران[5]، ۲۰۲۰)، (شجاع و همکاران[6]، ۲۰۲۰). در برخی از آخرین مطالعات انجام شده در حوزه تشخیص دیابت از روش‌های بیش نمونه‌گیری[7] جهت متوازن‌سازی داده‌های دیابت استفاده کرده‌اند که از جمله می‌توان به روش ارائه شده (رامش و همکاران[8]، ۲۰۲۱) اشاره کرد. این روش اگرچه مشکل عدم توازن داده‌ها را برطرف کرده است؛ اما به دلیل استفاده از ماشین بردار پشتیبان[9] که یک روش یادگیری ماشین سنتی است، در پایگاه داده دیابت هند با نام Pima به صحت[10] ۸۳ ٪ رسیده است که چندان موفق نیست. از این رو، ارائه یک روش تشخیص بیماری دیابت کارآمد مبتنی بر هوش‌مصنوعی و شبکه‌های عصبی عمیق بازگشتی است که در آن برای مشکل عدم توازن داده‌ها از الگوریتم بیش نمونه‌گیری SMOTE[11] یا نسخه‌های بهبود یافته آن مثل ADASYN[12] استفاده می‌شود، از جمله اهداف این پژوهش است. بخش‌های اصلی روش ارائه شده در این پژوهش به صورت زیر خلاصه می‌شود:

در این مدل از سه واحد پنهان بازگشتی مختلف شامل BiLSTM,[13]LSTM[14] و GRU[15] در شبکه عصبی بازگشتی استفاده شده است تا عملکرد هر یک مقایسه شود.

در این مدل سه الگوریتم بیش نمونه‌گیری مختلف شامل SMOTE، Borderline SMOTE و ADASYN جهت رفع مشکل داده‌های نامتوازن دیابت استفاده شده است و مقایسه بین عملکرد آن‌ها صورت گرفته است.

در ادامه این پژوهش، پیشینه پژوهش شامل مطالعات انجام شده در این زمینه و روش ارائه شده معرفی می‌شود و سپس در رابطه با نتایج حاصل شده، بحث می‌شود و در آخر نیز نتیجه‌گیری و کارهای آتی بیان خواهد شد.

مبانی نظری پژوهش

دیابت

دیابت در زمره بیماری‌های مزمن است که با افزایش سطح گلوکز خون مشخص می‌شود. در افراد مبتلا به این بیماری، عدم تولید انسولین توسط لوزالمعده یا عدم توانایی سلول‌ها در استفاده از انسولین، باعث افزایش گلوکز خون می‌شود. سه نوع اصلی دیابت وجود دارد: در دیابت نوع ۱، لوزالمعده قادر به تولید انسولین نیست. در نوع ۲، سلول‌های بدن در برابر انسولین مقاوم هستند و در طول زمان تولید انسولین کاهش پیدا می‌کند. در نوع ۳، دیابت حاملگی است که در دوره بارداری بروز می‌کند. افراد مبتلا به انواع دیابت در مقایسه با افرادی که مقادیر طبیعی گلوکز دارند، به شدت در معرض خطر ابتلا به بیماری‌های قلبی و عروقی قرار دارند (روگلیک[16]، ۲۰۱۶).

1. Fernández et al.
2. Resampling
3. Pertiwi et al.
4. Hairani et al.
5. Sreejith et al.
6. Shuja et al.
7. Over Sampling
8. Ramesh et al.
9. Support Vector Machine (SVM)
10. Accuracy
11. SMOTE: Synthetic Minority Over-sampling Technique
12. Adaptive Synthetic (ADASYN)
13. Bidirectional LSTM
14. Long Short-Term Memory
15. Gated Recurrent Unit
16. Roglic

دیابت نوع ۱ در کودکان، نوجوانان و جوانان بیشتر مشاهده می‌شود و هنوز علت آن مشخص نیست. پژوهشگران بر این باورند که ترکیبی از ژنتیک و عوامل محیطی منجر به دیابت نوع ۱ می‌شود. در مقابل عوامل خطر دیابت نوع ۲ بیشتر شناخته شده است. ژنتیک یکی از کلیدی‌ترین عوامل ابتلا به دیابت نوع ۲ است؛ اما در کنار آن اضافه وزن و چاقی و عدم تحرک بدنی از دیگر عوامل ابتلا به دیابت نوع ۲ است (روگلیک، ۲۰۱۶). همچنین نتایج مطالعات نشان می‌دهد که سیگار کشیدن نیز از عوامل افزایش احتمال ابتلا به دیابت است؛ اما در بین عوامل فوق قوی‌ترین عامل خطر، افزایش چربی بدن است. عادات غذایی غلط مانند مصرف زیاد قند و چربی نیز با افزایش خطر ابتلا به دیابت نوع ۲ ارتباط مستقیم دارد. در خطر دیابت بارداری، سابقه خانوادگی، سن، اضافه وزن و چاقی، بی‌تحرکی فیزیکی و اضافه وزن بیش از حد در دوران بارداری از علل اصلی بروز این بیماری است. دیابت کنترل نشده، عوارض بسیاری در بدن بیمار ایجاد می‌کند. این بیماری باعث آسیب به رگ‌های خونی و اعصاب می‌شود و در نتیجه، از دست دادن بینایی و عملکرد کلیه، حملات قلبی، سکته مغزی و قطع اندام تحتانی رخ می‌دهد. دیابت از عوامل ناتوانی و کوتاهی عمر است (روگلیک، ۲۰۱۶).

باز نمونه‌گیری: بیش نمونه‌گیری و کم نمونه‌گیری

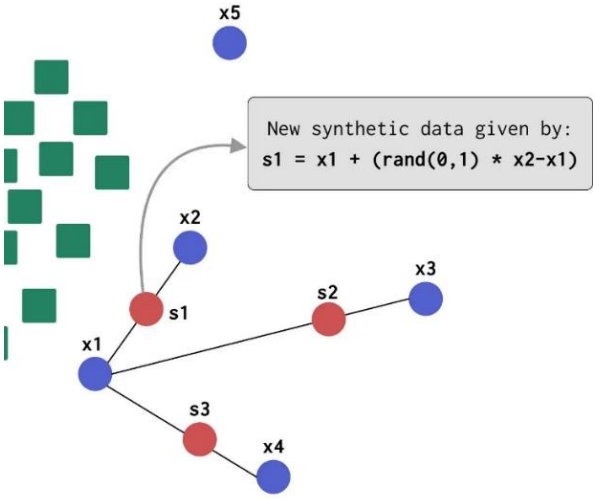
عدم توازن داده‌ها باعث می‌شود، روش‌های یادگیری ماشین عملکرد ضعیفی از خود ارائه دهند. در داده‌های نامتوازن یک یا چند کلاس در مقایسه با کلاس‌های دیگر دارای داده‌های بسیار کمتری هستند که به آن داده‌های اقلیت گفته می‌شود. داده‌های اقلیت در یادگیری مدل طبقه‌بندی کننده بی‌تأثیر هستند، در نتیجه مدل نمی‌تواند کلاس آن‌ها را به درستی پیش‌بینی کند. برای کنترل داده‌های نامتوازن روش‌های بازنمونه‌گیری مطرح شدند. روش‌های باز نمونه‌گیری به صورت کم نمونه‌گیری و یا بیش نمونه‌گیری هستند.

روش‌های بیش نمونه‌گیری با توجه به همسایه‌های داده‌های کلاس اقلیت، داده‌های مصنوعی تولید می‌کنند تا داده‌های کلاس اقلیت افزایش پیدا کند. تولید داده‌های مصنوعی دقیق که شباهت زیادی به داده اقلیت اصلی داشته باشد در این الگوریتم‌ها بسیار مهم است؛ چرا که اگر داده‌های مصنوعی با دقت کم تولید شوند در نهایت دقت طبقه‌بندی نیز کاهش پیدا می‌کند.

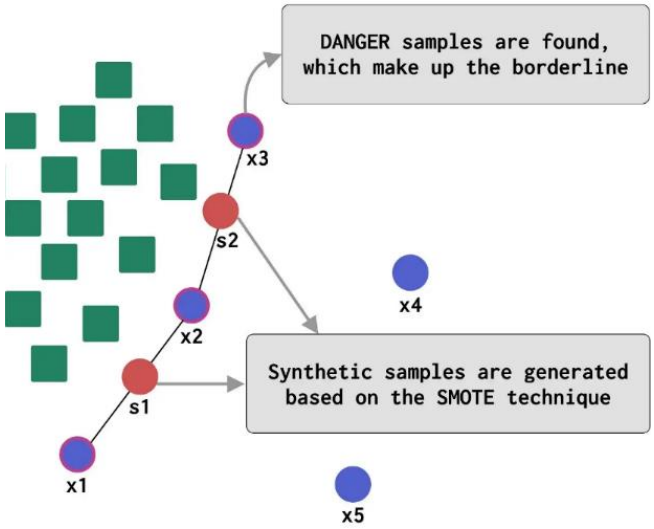
الگوریتم بیش نمونه‌گیری SMOTE در سال ۲۰۰۲ توسط چاولا و همکاران[1] معرفی شد. در این روش برای هر داده از کلاس اقلیت، k نزدیک‌ترین همسایه‌های آن جستجو می‌شود و برخی از همسایه‌ها به طور تصادفی با توجه به نرخ بیش نمونه‌گیری انتخاب می‌شوند. سپس، داده‌های مصنوعی جدید در امتداد خط، بین داده اقلیت و نزدیک‌ترین همسایه‌های منتخب تولید می‌شوند. این الگوریتم از کل داده‌های کلاس اقلیت استفاده می‌کند (چاولا و همکاران، ۲۰۰۲). در شکل ۱ عملکرد این الگوریتم قابل مشاهده است.

الگوریتم بیش نمونه‌گیری Borderline SMOTE در سال ۲۰۰۵ توسط هان و همکاران[2] معرفی شد. این الگوریتم برای تولید داده‌های مصنوعی، از داده‌های مرزی استفاده می‌کند. داده‌های مرزی نسبت به داده‌هایی که دور از مرز هستند، مستعد طبقه‌بندی اشتباه هستند؛ در نتیجه برای طبقه‌بندی، اهمیت بیشتری دارند. در این روش بر خلاف SMOTE فقط داده‌های اقلیت مرزی بیش نمونه‌گیری می‌شوند. در این روش ابتدا، داده‌های اقلیت مرزی جستجو می‌شوند، سپس داده‌های مصنوعی با توجه به آن‌ها تولید شده و به مجموعه اصلی اضافه می‌شود (هان و همکاران ۲۰۰۵). در شکل ۲ عملکرد این مدل نشان داده شده است.

1. Chawla et al.
2. Han et al.



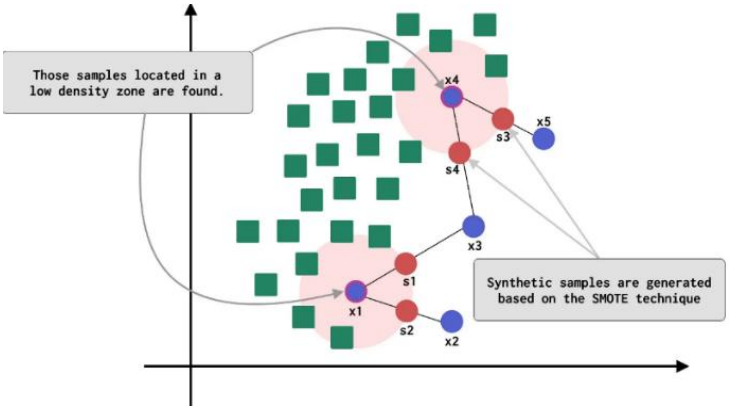
شکل ۱: بیش نمونه‌گیری به روش SMOTE (لوپز، ۲۰۲۱)



شکل ۲: بیش نمونه‌گیری به روش Borderline (لوپز[1]، ۲۰۲۱)

روش بیش نمونه‌گیری تطبیقی یا به اختصار (ADASYN) توسط هایبو و همکاران[2] در سال ۲۰۰۸ معرفی شد. ADASYN بر اساس الگوریتم SMOTE برای تولید داده‌های مصنوعی بکار می‌رود. تفاوت بین ADASYN و SMOTE در این است که روش تطبیقی، نمونه‌های کلاس اقلیت را به نحوی شناسایی می‌کند که تولید نمونه در مناطق با چگالی پایین کلاس اقلیت انجام شود؛ یعنی، روش تطبیقی بر روی نمونه‌هایی از کلاس اقلیت تمرکز می‌کند که طبقه‌بندی آن‌ها دشوار است؛ زیرا در یک منطقه کم تراکم قرار دارند (هایبو و همکاران، ۲۰۰۸). در شکل ۳ عملکرد این روش قابل مشاهده است.

1. López.
2. Haibo et al.



شکل ۳: بیش نمونه‌گیری به روش تطبیقی (لوپز، ۲۰۲۱)

شبکه عصبی عمیق بازگشتی

شبکه‌های عصبی بازگشتی، به طور گسترده در داده‌های توالی، مانند متن، صدا و ویدئو استفاده می‌شوند. با این حال، شبکه‌های بازگشتی اولیه از سلول‌های سیگما یا سلول‌های تانژانت هایپربولیک تشکیل شده‌اند و یکی از نقاط ضعف اصلی آن‌ها این است که در شرایطی که شکاف توالی ورودی زیاد است، قادر به یادگیری اطلاعات مربوط به داده‌های ورودی نیستند (یو و همکاران[1]، ۲۰۱۹). در رابطه ۱ عملکرد سلول سیگما بازگشتی استاندارد بیان شده است.

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b), \qquad y_t = h_t \tag{۱}$$

در رابطه بالا x_t، h_t و y_t به ترتیب نماد ورودی، اطلاعات بازگشتی و خروجی سلول در زمان t هستند. W_h و W_x وزن و b نماد مقدار بایاس[2] است (یو و همکاران، ۲۰۱۹). همان‌طور که بیان شد، شبکه‌های بازگشتی اولیه که از سلول‌های بازگشتی استاندارد (رابطه۱) تشکیل شده‌اند، قادر به مدیریت وابستگی‌های طولانی مدت نیستند. برای رفع این چالش در سال ۱۹۹۷ هوکرایتر و همکارانش[3] با افزودن توابع دروازه به ساختار سلول بازگشتی اولیه، واحد پنهان جدیدی با عنوان حافظه بلندمدت کوتاه (LSTM) معرفی کردند که توانست مشکل وابستگی‌های بلندمدت را به خوبی حل کند (یو و همکاران ۲۰۱۹). آن‌ها در سلول جدید، ظرفیت یادآوری[4] سلول بازگشتی استاندارد را بهبود بخشیدند؛ عملیات انجام شده در این سلول جدید در رابطه ۲ بیان شده است (یو و همکاران ۲۰۱۹).

$$\begin{aligned} i_t &= \sigma(W_{ih} h_{t-1} + W_{ix} x_t + b_i), \\ \bar{c_t} &= (W_{\bar{c}h} h_{t-1} + W_{\bar{c}x} x_t + b_{\bar{c}}), \\ c_t &= c_{t-1} + i_t . \bar{c}_t, \\ o_t &= (W_{oh} h_{t-1} + W_{ox} x_t + b_o), \\ h_t &= o_t . \tanh(c_t). \end{aligned} \tag{۲}$$

در رابطه بالا c_t نماد وضعیت سلول[5]، W نماد ماتریس‌های وزنی و $(.)$ نماد عملگر ضرب نقطه‌ای[6] است. در فرآیند به‌روزرسانی وضعیت سلول، دروازه ورودی تصمیم می‌گیرد که چه اطلاعات جدیدی در سلول ذخیره شود و دروازه خروجی تصمیم می‌گیرد که چه اطلاعاتی بر اساس وضعیت سلول به عنوان خروجی تولید شود (یو و همکاران، ۲۰۱۹). محققان زیادی در راستای بهبود سلول LSTM تلاش کردند. در سال ۲۰۰۰ گرس و همکارانش[7] دروازه فراموشی[8] را

1. Yu et al.
2. Bias
3. Hochreiter et al.
4. Remembering capacity
5. Cell State
6. Pointwise Multiplication
7. Gers et al.
8. Forget gate

به سلول LSTM افزودند که عملیات درونی آن در رابطه ۳ ارائه شده است (گرس و همکاران، ۲۰۰۰). دروازه فراموشی تصمیم می‌گیرد که چه اطلاعاتی از وضعیت سلول حذف شود. هنگامی که مقدار دروازه فراموشی f_t، ۱ است، اطلاعات حفظ می‌شود، در غیر این صورت تمام اطلاعات حذف می‌شود (یو و همکاران، ۲۰۱۹).

$$
\begin{aligned}
f_t &= \sigma\left(W_{fh}h_{t-1} + W_{fx}x_t + b_f\right)\\
i_t &= \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i),\\
\bar{c_t} &= (W_{\bar{c}h}h_{t-1} + W_{\bar{c}x}x_t + b_{\bar{c}}),\\
c_t &= c_{t-1} + i_t.\bar{c}_t,\\
o_t &= (W_{oh}h_{t-1} + W_{ox}x_t + b_o),\\
h_t &= o_t.\tanh(c_t).
\end{aligned}
\qquad (۳)
$$

سلول‌های بازگشتی BiLSTM برای تحلیل وابستگی‌های دوطرفه از داده‌های توالی معرفی شدند. در این سلول‌ها بهترین توالی‌ها بر اساس توالی‌های قبلی و بعدی مشخص می‌شود. توالی ورودی اول به لایه بازگشتی اول ارائه می‌شود و معکوس آن به لایه بازگشتی دوم ارائه می‌شود. دو لایه اول و دوم به لایه خروجی مشترک متصل می‌شوند (کوی و همکاران[۱]، ۲۰۱۸).

سلول‌های بازگشتی GRU در سال ۲۰۱۴ معرفی شدند و مبتنی بر سلول LSTM هستند؛ اما پارامترهای کمتری در مقایسه با LSTM دارند؛ از این رو مرحله یادگیری آن‌ها سریع‌تر است. این سلول‌ها در مقابل چالش محوشدگی گرادیان[۲] مقاوم هستند. واحد پنهان GRU دارای دو دروازه به‌روزرسانی[۳] و تنظیم مجدد[۴] است.

پیشینهٔ پژوهش

مطالعات زیادی بین سال‌های ۲۰۱۰ تا ۲۰۱۸ برای تشخیص دیابت انجام شده است که اغلب این روش‌ها مبتنی بر روش‌های یادگیری ماشین سنتی مثل درخت تصمیم (کنداسمی و بارامورالی[۵]، ۲۰۱۵)، (منگ و همکاران[۶]، ۲۰۱۳)، (نیلاشی و همکاران[۷]، ۲۰۱۷)، (پروین و همکاران[۸]، ۲۰۱۶)، (آکسو و همکاران[۹]، ۲۰۱۷)، K نزدیک‌ترین همسایه (ساکسنا و همکاران[۱۰]، ۲۰۱۴) بیز ساده (گارگا و همکاران[۱۱]، ۲۰۱۷)، (کومار دوانگان و آگراوال[۱۲]، ۲۰۱۵)، ماشین بردار پشتیبان (سانتنام و پادماواتی[۱۳]، ۲۰۱۵) بودند.

(داس و همکاران[۱۴]، ۲۰۱۸) از فناوری‌های درخت تصمیم J48 و بیز ساده برای تشخیص دیابت استفاده کردند. این تحقیق روشی سریع و کارآمد برای تشخیص دیابت ارائه داده است. آن‌ها بر روی پایگاه داده جمع‌آوری شده با الگوریتم درخت تصمیم به صحت ۶۹/۵٪ و با الگوریتم بیز ساده به صحت ۷۲/۵٪ دست پیدا کردند.

1. Cui et al.
2. Vanishing Gradient
3. Update
4. Reset
5. Kandhasamy & Balamurali
6. Meng et al.
7. Nilashi et al.
8. Perveen et al.
9. Xu et al.
10. Saxena et al.
11. Garga et al.
12. kumar Dewangan & Agrawal
13. Santhanam & Padmavathi
14. Das et al.

(وو و همکاران[1]، ۲۰۱۸) یک مدل ترکیبی ارائه دادند. آن‌ها از الگوریتم خوشه‌بندی بهبود یافته K-Means و الگوریتم رگرسیون لجستیک استفاده کردند که به دقت بالاتری رسیدند. روش آن‌ها در پایگاه داده دیابت Pima به صحت ۹۵/۴۱٪ رسیده است.

(سیسودیا و همکاران[2]، ۲۰۱۸) از سه الگوریتم درخت تصمیم، ماشین بردار پشتیبان و بیز ساده برای تشخیص دیابت در مراحل اولیه استفاده کردند. در بین الگوریتم‌ها، الگوریتم بیز ساده بر روی پایگاه داده PIDD به صحت ۷۶/۳۰٪ رسید. کمترین صحت در بین سه روش فوق ۶۵/۱۰٪ است که متعلق به ماشین بردار پشتیبان است.

(کاناداسان و همکاران[3]، ۲۰۱۹) برای طبقه‌بندی دیابت در پایگاه داده Pima یک روش مبتنی بر شبکه عصبی عمیق ارائه دادند. در این روش از خودرمزگذارها[4] برای استخراج ویژگی استفاده و سپس با استفاده از همان معماری تشخیص دیابت انجام شد. روش آن‌ها بر روی پایگاه داده Pima به صحت ۸۶/۲۶٪ رسیده است.

(رانجی و همکاران[5]، ۲۰۱۹) جهت تشخیص دیابت مدلی مبتنی بر بیز ساده به نام RB-Bayes در داده‌های Pima ارائه کردند. در این روش ابتدا مشکل داده‌های از دست رفته برطرف شده و سپس روش RB-Bayes بر روی داده‌های پیش‌پردازش شده اعمال می‌شود. این روش به صحت ۷۲/۹٪ دست یافته است.

(ونگ و همکاران[6]، ۲۰۱۹) روشی برای طبقه‌بندی داده‌های Pima ارائه کردند که در آن به مشکل داده‌های نامتوازن و داده‌های از دست رفته پرداخته‌اند. در این روش با استفاده از الگوریتم بیز ساده مشکل داده‌های از دست رفته برطرف شده و سپس با استفاده از الگوریتم بیش نمونه‌گیری ADASYN مشکل داده‌های نامتوازن کنترل شده است. با استفاده از الگوریتم جنگل تصادفی تشخیص دیابت انجام شده است. این روش به صحت ۸۷/۱٪ رسیده است.

(هوما و همکاران[7]، ۲۰۲۰) یک مطالعه مقایسه‌ای انجام دادند که در آن از شبکه‌های عصبی، بیز ساده، درخت تصمیم و شبکه‌عصبی پرسپترون چندلایه برای تشخیص دیابت در پایگاه داده Pima استفاده کردند. در این روش بهترین صحت در داده‌های Pima برابر با ۸۹/۰۷٪ است که توسط شبکه‌عصبی پرسپترون چندلایه‌ای حاصل شده و کمترین صحت برای الگوریتم بیز ساده است که برابر با ۷۶/۳۳٪ است.

(عابدینی و همکاران[8]، ۲۰۲۰) یک مدل سلسله مراتبی برای ترکیب دو یا چند روش طبقه‌بندی پیشنهاد دادند. ابتدا یک درخت تصمیم و یک مدل رگرسیون لجستیک آموزش داده شده و سپس در مرحله دوم خروجی این مدل‌ها به شبکه عصبی ارائه شده است. مدل پیشنهاد شده توسط آن‌ها بر روی پایگاه داده Pima به صحت ۸۳/۰۸ ٪ رسیده است.

(چوبی و همکاران[9]، ۲۰۲۰) از چندین الگوریتم مختلف شامل Adaboost، رگرسیون خطی، شبکه تابع پایه شعاعی و نزدیک‌ترین همسایه برای تشخیص دیابت در پایگاه داده Pima استفاده کردند. در این روش از دو الگوریتم کاهش ابعاد تحلیل مؤلفه اصلی و تجزیه‌وتحلیل متمایز خطی برای کاهش ابعاد داده استفاده شده است. بهترین نتیجه متعلق به ترکیب الگوریتم تحلیل مؤلفه اصلی و رگرسیون خطی است.

(ناموکو و همکاران[10]، ۲۰۲۰) از فناوری بیش نمونه‌گیری SMOTE برای رفع مشکل داده‌های نامتوازن در پایگاه داده Pima استفاده کردند و سپس با استفاده از الگوریتم‌های ماشین بردار پشتیبان، بیز ساده و درخت تصمیم C4.5 تشخیص دیابت را انجام دادند. در این روش بهترین صحت برابر با ۸۹/۵٪ است که توسط الگوریتم درخت تصمیم C4.5 به دست

1. Wu et al.
2. Sisodia et al.
3. Kannadasan et al.
4. stacked autoencoders
5. Rajni et al.
6. Wang et al.
7. Huma et al.
8. Abedini et all.
9. Choubey et al.
10. Nnamoko et al.

آمده است؛ اگرچه در این روش از چندین الگوریتم برای تشخیص دیابت استفاده شده است؛ اما نتایج گروهی آن‌ها بررسی نشده است.

(رامش و همکارانش، ۲۰۲۱) یک روش مبتنی بر مدل‌های یادگیری ماشین سنتی برای تشخیص دیابت ارائه دادند. این روش اگرچه مشکل عدم توازن داده‌ها را با استفاده از الگوریتم SMOTE برطرف کرده است؛ اما به دلیل استفاده از ماشین بردار پشتیبان[1] که یک روش یادگیری ماشین سنتی است، در پایگاه داده دیابت هند با نام Pima به صحت ۸۳٪ رسیده است که چندان موفق نیست.

(چنگ و همکارانش[2]، ۲۰۲۲) نیز از روش‌های یادگیری ماشین سنتی برای تشخیص دیابت در پایگاه داده Pima استفاده کردند. در این مدل ابتدا پیش‌پردازش بر روی داده‌ها انجام شده است و سپس انتخاب ویژگی صورت گرفته و با استفاده از سه الگوریتم جنگل تصادفی و درخت تصمیم J48 و بیز ساده تشخیص دیابت انجام شده است. در این مدل از خوشه‌بندی k-means و الگوریتم تحلیل مؤلفه اصلی[3] برای انتخاب ویژگی استفاده شده است. این مدل در شرایطی که ۵ ویژگی را انتخاب کرده است با استفاده از بیز ساده به صحت ۷۷/۸۳٪ و دقت ۸۱/۲۵٪ دست یافته است.

جدول ۱: خلاصه پیشینه پژوهش

نویسنده/ سال	عملکرد	پایگاه داده	نتایج	مزایا/معایب
منگ و همکاران ۲۰۱۳	شبکه‌های عصبی مصنوعی، رگرسیون لجستیک و مدل درخت تصمیمC5.0	گوانگژو چین	صحت ۷۷/۸۷٪	مزایا: مقایسه چندین الگوریتم معایب: صحت پایین
کاراتی و همکاران ۲۰۱۴	الگوریتم k نزدیک‌ترین همسایه	LARS	صحت ۷۰٪	مزایا: سادگی روش معایب: صحت پایین، وابستگی نتایج به مقدار K
بالامورالی و همکاران ۲۰۱۵	درخت تصمیم j48، k نزدیک‌ترین همسایه و جنگل تصادفی و ماشین بردار پشتیبان	Pima	صحت ۱۰۰٪	مزایا: بررسی و مقایسه چندین الگوریتم. حصول بهترین صحت و توجه به داده‌های پرت و رفع مشکل داده‌های پرت معایب: عدم رفع مشکل داده‌های نامتوازن
اگراوال و همکاران ۲۰۱۵	بیز ساده و شبکه عصبی چندلایه پرسپترون	Pima	صحت برابر با ۸۱/۸۹٪	مزایا: ترکیب روش‌های یادگیری ماشین و شبکه عصبی معایب: مشکل داده‌های نامتوازن و ناقص رفع نشده است
سانتانام و همکاران ۲۰۱۵	خوشه‌بندی K-Means، فرااکتشافی ژنتیک و ماشین بردار پشتیبان	Pima	صحت ۹۸/۷۸٪	مزایا: رفع مشکل داده‌های پرت و صحت بالا معایب: ماهیت تصادفی نتایج به دلیل مرحله انتخاب ویژگی مبتنی بر ژنتیک
پروین و همکاران ۲۰۱۶	Adaboost و Bagging با استفاده از J48 Decision Tree	CPCSSN	AROC برابر با ۰/۹۸٪	مزایا: استفاده از الگوریتم گروهی معایب: عدم بررسی صحت. عدم استفاده از پایگاه داده عمومی
گارگ و همکاران ۲۰۱۷	شامل بیز ساده، شبکه بیزی، درخت تصمیم J48، طبقه‌بندی (SMO)، جنگل تصادفی	Pima	صحت ۷۷/۳۴٪	مزایا: مقایسه روش‌های مختلف معایب: صحت پایین و وابستگی نتایج به تابع هسته در SMO

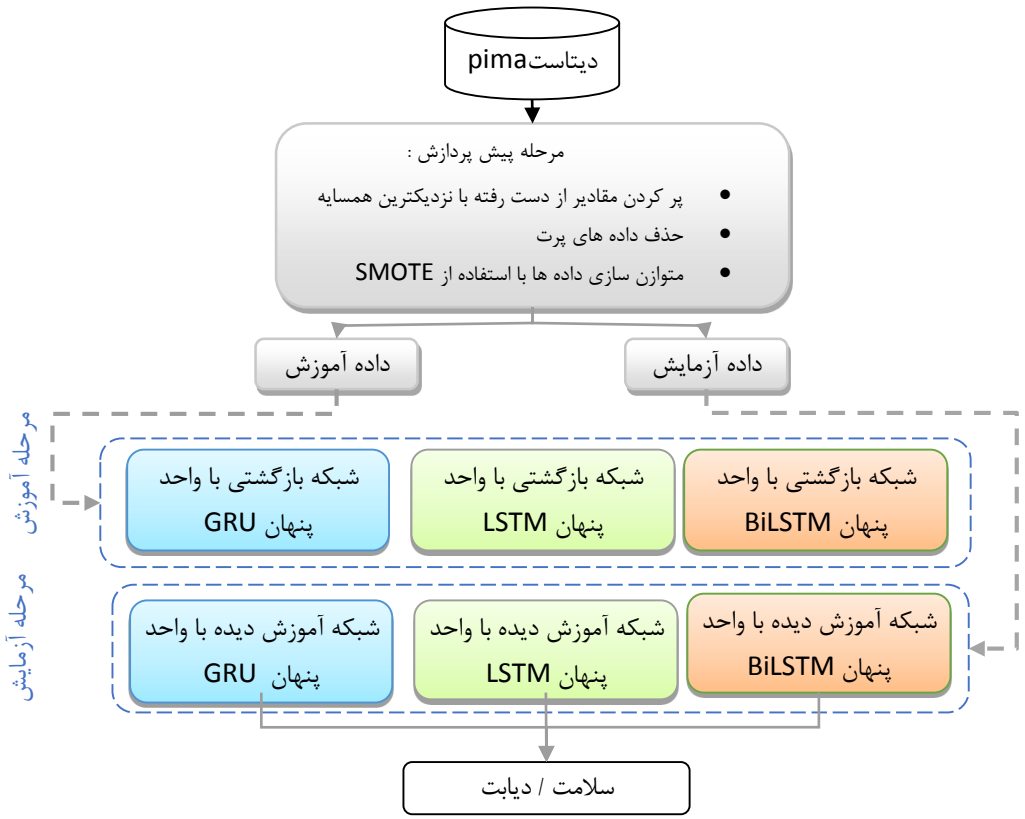
1. Support Vector Machine (SVM)
2. Chang et al.
3. k-means clustering, principal component analysis (PCA)

نویسنده/ سال	عملکرد	پایگاه داده	نتایج	مزایا/معایب
خو و همکاران ۲۰۱۷	جنگل تصادفی	Pima	صحت ٪۸۵	مزایا: سادگی روش معایب: استفاده از یک الگوریتم یادگیری، عدم حذف داده‌های گمشده
نیلاشی و همکاران ۲۰۱۷	درخت تصمیم CART، خوشه‌بندی EM، تکنیک تحلیل مؤلفه اصلی	Pima	صحت ٪۹۲/۹	مزایا: رفع مشکل داده‌های پرت و ترکیب خوشه‌بندی و طبقه‌بندی معایب: عدم رفع داده‌های نامتوازن
داس و همکاران ۲۰۱۸	درخت تصمیم J48 و بیز ساده	جمع‌آوری شده	صحت ٪۷۲/۵	مزایا: سرعت بالا، مقایسه دو الگوریتم معایب: صحت پایین
وو و همکاران ۲۰۱۸	خوشه‌بندی بهبود یافته K-Means و الگوریتم رگرسیون لجستیک	Pima	صحت ٪۹۵/۴۱	مزایا: استفاده از روش ترکیبی معایب: مشکل در تعیین تعداد خوشه بهینه و کاهش صحت در صورت عدم تخصیص بهینه تعداد خوشه
سیسودیا و همکاران ۲۰۱۸	درخت تصمیم، الگوریتم ماشین بردار پشتیبان و بیز ساده	PIDD	صحت ٪۷۶/۳۰	مزایا: بررسی سه روش یادگیری ماشین معایب: صحت پایین. عدم ترکیب روش‌ها برای رسیدن به نتایج بالاتر
کاناداسان و همکاران ۲۰۱۹	شبکه عصبی عمیق خودرمزگذارها	Pima	صحت ٪۸۶/۲۶	مزایا: استفاده از معماری عمیق کارآمدی روش در داده‌های بدون ناظر معایب: کاهش نتایج در صورت تعریف نامناسب نودهای رمزگذار
رانجی و همکاران ۲۰۱۹	بر بیز ساده به نام RB-Bayes	Pima	صحت ٪۷۲/۹	مزایا: رفع مشکل داده‌های از دست رفته معایب: صحت پایین
ونگ و همکاران ۲۰۱۹	بیز ساده، بیش نمونه‌گیری ADASYN، جنگل تصادفی	Pima	صحت ٪۸۷/۱	مزایا: رفع مشکل داده‌های نامتوازن و از دست رفته معایب: استفاده از یک الگوریتم یادگیری ماشین به جای روش‌های گروهی
هوما و همکاران ۲۰۲۰	شبکه‌های عصبی، بیز ساده، درخت تصمیم و شبکه عصبی پرسپترون چندلایه	Pima	صحت ٪۸۹/۰۷	مزایا: مقایسه روش‌های یادگیری ماشین و شبکه عصبی معایب: عدم پیش‌پردازش داده‌ها
عابدینی و همکاران ۲۰۲۰	مدل سلسله مراتبی درخت تصمیم و یک مدل رگرسیون لجستیک	Pima	صحت ٪۸۳/۰۸	مزایا: تعریف یک مدل ترکیبی دوسطحی معایب: عدم رفع مشکل داده‌های نامتوازن
چوبی و همکاران ۲۰۲۰	Adaboost، رگرسیون خطی، شبکه تابع پایه شعاعی و نزدیک‌ترین همسایه	Pima	-	مزایا: کاهش ابعاد معایب: عدم پیش‌پردازش داده‌ها
ناموکو همکاران ۲۰۲۰	بیش نمونه‌گیری SMOTE، ماشین بردار پشتیبان، بیز ساده و درخت تصمیم C4.5	Pima	صحت ٪۸۹/۵	مزایا: رفع چالش داده‌های نامتوازن معایب: استفاده از یک الگوریتم طبقه‌بندی، عدم رفع مشکل داده‌های از دست رفته، استفاده از نسخه ساده SMOTE

نویسنده/ سال	عملکرد	پایگاه داده	نتایج	مزایا/معایب
رامش و همکارانش ۲۰۲۱	بیش نمونه‌گیری SMOTE، ماشین بردار پشتیبان	Pima	صحت ٪۸۳	معایب: استفاده از یک الگوریتم طبقه‌بندی، استفاده از نسخه ساده SMOTE
چنگ و همکارانش ۲۰۲۲	مقداردهی به داده‌های از دست رفته و انتخاب ویژگی با PCA و K-means	Pima	صحت: ٪۷۷,۸۳	مزایا: بررسی چند الگوریتم مختلف معایب: عدم بررسی روش‌های گروهی و ترکیبی، عدم توجه به داده‌های نامتوازن

روش پژوهش

مدل ارائه شده در این مقاله یک روش طبقه‌بندی مبتنی بر یادگیری عمیق است که جهت تشخیص دیابت از سه شبکه عصبی عمیق بازگشتی با سه واحد پنهان مختلف استفاده می‌کند. هدف از انتخاب سه شبکه بازگشتی بررسی عملکرد آن‌ها در تشخیص دیابت و یافتن بهترین معماری برای این هدف است. دیاگرام روش ارائه شده در شکل ۴ ارائه شده است. این مدل شامل سه مرحله از جمله ۱) آماده‌سازی داده‌ها، ۲) طبقه‌بندی، ۳) ارزیابی است. گام آماده‌سازی خود از سه مرحله مقداردهی داده‌های گمشده، حذف داده‌های پرت و بیش نمونه‌گیری تشکیل شده است.



شکل ۴: مدل پیشنهادی

مقداردهی مقادیر از دست رفته و حذف داده‌های پرت

مرحله آماده‌سازی داده‌ها در روش‌های یادگیری ماشین از اهمیت بسزایی برخوردار است؛ چرا که داده‌های با کیفیت پایین به شدت منجر به کاهش عملکرد مدل‌های یادگیری ماشین می‌شود. این گام شامل سه زیر گام مقداردهی به داده‌های از دست رفته و حذف داده‌های پرت و متوازن‌سازی داده‌ها است. برای مقداردهی داده‌های از دست رفته، روش‌های زیادی وجود دارد، به عنوان مثال برخی از روش‌ها داده‌هایی که دارای مقادیر گمشده هستند را نادیده می‌گیرند؛ این روش‌ها باعث می‌شوند، اطلاعات مفید زیادی از دست برود و کماکان منجر به کاهش عملکرد روش‌های طبقه‌بندی می‌شوند. در برخی دیگر از مدل‌ها، مقدار گمشده به صورت دستی پر می‌شوند. این رویکرد عملاً در داده‌های بزرگ کاربردی نیست و فرآیندی بسیار هزینه‌بر و زمان‌بر است. در روشی دیگر مقادیر از دست رفته با یک مقدار ثابت سراسری یا میانگین داده‌ها جایگزین می‌شود. این رویکرد بر این فرض استوار است که تمام مقادیر گمشده دارای یک مقدار هستند، این روش کنترل مقادیر از دست رفته منجر به تحریفات بزرگی در توزیع داده‌ها می‌شود. آخرین رویکرد که در این پژوهش نیز استفاده شده است، محتمل‌ترین مقدار را برای پر کردن مقادیر گمشده محاسبه می‌کند. این رویکرد، از اطلاعات مشاهده شده برای پر کردن مقادیر گمشده استفاده می‌کند.

در این مدل از رویکرد چهارم و از الگوریتم نزدیک‌ترین همسایه که یک روش غیرپارامتری است برای پر کردن مقادیر گمشده استفاده شده است. روش‌های غیرپارامتری از این جهت مناسب هستند که خروجی و عملکرد آن‌ها وابسته به پارامترهای ورودی نیست، در نتیجه نتایج قابل اطمینانی تولید می‌کنند. عملکرد این الگوریتم برای رفع این چالش به این صورت است که ابتدا نزدیک‌ترین همسایه‌ها به هر داده‌ای که دارای مقدار از دست رفته است را با استفاده از فاصله اقلیدسی پیدا می‌کند. سپس اولین همسایه که دارای کمترین فاصله یا بیشترین شباهت به داده ناقص است را انتخاب کرده و مقدار متناظر در داده کامل را با داده ناقص جایگزین می‌کند. فاصله اقلیدسی از رابطه ۴ محاسبه می‌شود. عملکرد الگوریتم نزدیک‌ترین همسایه برای این بخش از مدل با ذکر یک مثال نشان داده شده است.

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (۴)$$

در رابطه (۴) x و y دو نمونه از داده‌ها هستند و n ابعاد داده‌ها است.

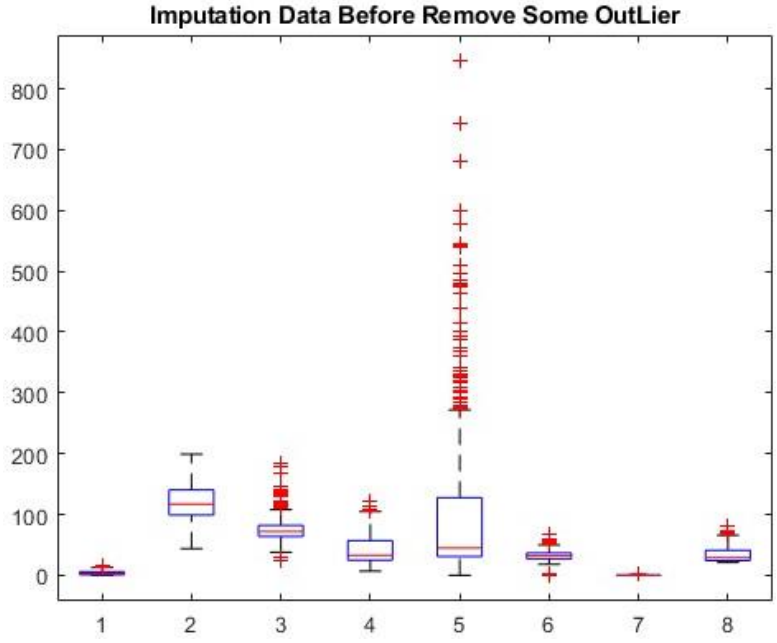
اگر ماتریس فرضی A دارای مقادیری به صورت زیر باشد و ستون اول در آن دارای مقدار گمشده باشد و نزدیک‌ترین همسایه به ستون اول، با استفاده از نزدیک‌ترین همسایه و فاصله اقلیدسی، ستون دوم باشد؛ آنگاه در مقدار گمشده ستون اول، با توجه به مقدار درایه متناظر در ستون دوم، مقدار ۱- درج می‌شود.

$$A = \begin{bmatrix} 1 & 2 & 5 \\ 4 & 5 & 7 \\ NAN & -1 & 8 \\ 7 & 6 & 0 \end{bmatrix} = impute\ using\ NN = A(3,1) = -1 \rightarrow\rightarrow\rightarrow A = \begin{bmatrix} 1 & 2 & 5 \\ 4 & 5 & 7 \\ -1 & -1 & 8 \\ 7 & 6 & 0 \end{bmatrix}$$

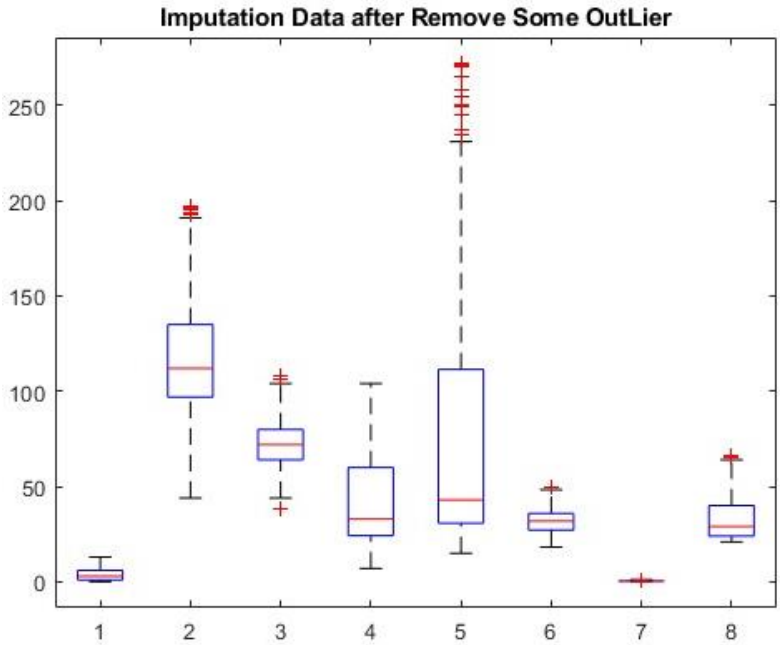
در این مدل، برای حذف داده‌های پرت از یک روش مبتنی بر دامنه یعنی نمودار چارک[1]، استفاده شد. نمودار چارک داده‌های دیابت این مطالعه، قبل و بعد از حذف داده‌های پرت در شکل ۵ و ۶ نشان داده شده است. این نمودار موقعیت و پراکندگی داده‌ها را به خوبی نشان می‌دهد. این نمودار با استفاده از مستطیل‌ها و نقطه‌چین دو طرف آن‌ها و با استفاده از مقادیر میانه، حداقل و حداکثر، چارک‌های اول و سوم ترسیم می‌شود. در این نمودارها طول مستطیل‌ها نشان‌دهنده فاصله چارکی است. در واقع طول هر یک از مستطیل‌ها فاصله چارک اول از سوم را نشان می‌دهد و از رابطه $IQR = Q_3 - Q_1$ حاصل می‌شود. در این فناوری، داده‌های پرت با توجه به دو شرط مشخص شده و حذف می‌شوند. داده‌هایی

1. Quartiles

که کوچک‌تر از $Q_1 - 1.5IQR$ و یا بزرگ‌تر از $Q_3 + 1.5IQR$ باشند، داده پرت ضعیف و داده‌هایی که کوچک‌تر از $Q_1 - 3IQR$ و یا بزرگ‌تر از $Q_3 + 3IQR$ باشند، داده پرت قوی هستند و حذف می‌شوند.



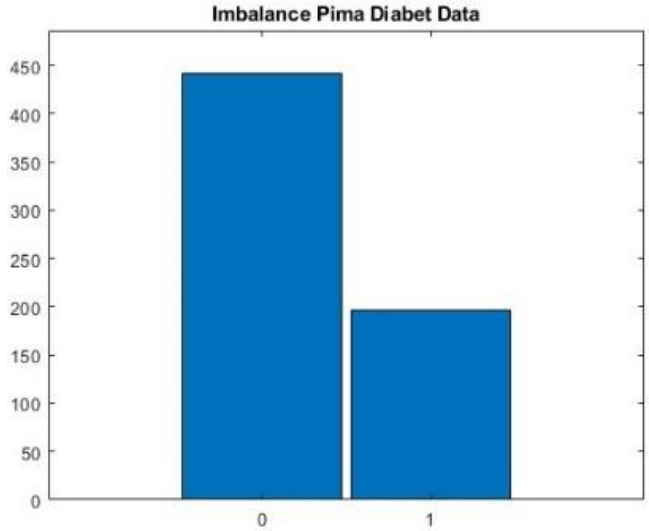
شکل ۵: نمودار چارکی داده‌ها قبل از حذف داده‌های پرت



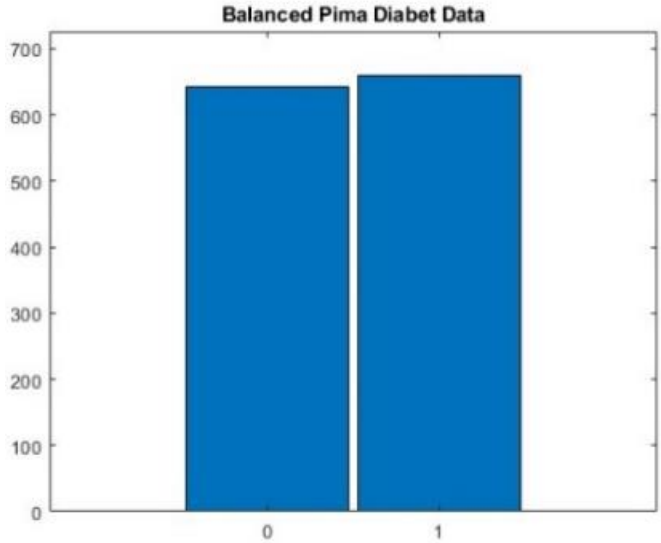
شکل ۶: نمودار چارکی داده‌ها بعد از حذف داده‌های پرت

متوازن‌سازی داده‌ها با بیش نمونه‌گیری

داده‌های حاصل شده از دو مرحله قبل به الگوریتم SMOTE، Borderline SMOTE و ADASYN ارائه می‌شود تا هر یک از این الگوریتم‌ها به شیوه خود داده‌های مصنوعی را تولید کرده و به کلاس اقلیت اضافه کنند. در این پژوهش تعداد همسایگان برای هر سه الگوریتم بیش نمونه‌گیری برابر با ۲ در نظر گرفته شد. در این حالت هر یک از سه الگوریتم مذکور برای هر داده در کلاس اقلیت، دو همسایه پیدا کرده و با استفاده از داده‌های دو همسایه داده مصنوعی جدید را تولید می‌کنند. در شکل ۷ و ۸ نمودار فراوانی دو کلاس دیابت و سالم بعد و قبل از متوازن‌سازی داده‌ها ترسیم شده است.



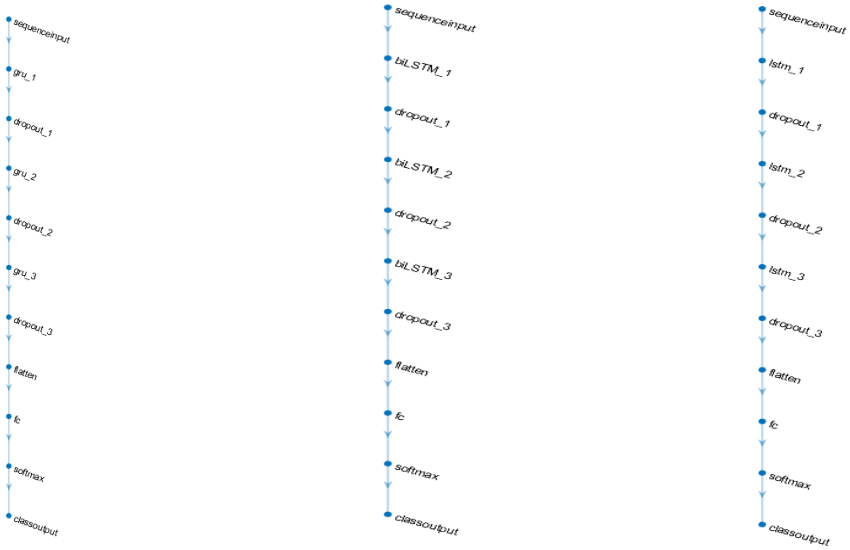
شکل ۷: تعداد داده‌ها در دو کلاس سالم و دیابت قبل از متوازن‌سازی به روش بیش نمونه‌گیری



شکل ۸: تعداد داده‌ها در دو کلاس سالم و دیابت بعد از متوازن‌سازی به روش بیش نمونه‌گیری

تشخیص دیابت با سه واحد پنهان بازگشتی

معماری تعریف شده برای سه شبکه در شکل ۹ ترسیم شده است. سه معماری تعریف شده با یکدیگر مشابه هستند و فقط در نوع واحد پنهان با هم متفاوت هستند.



معماری شبکه عصبی بازگشتی با واحد پنهان LSTM | معماری شبکه عصبی بازگشتی با واحد پنهان BiLSTM | معماری شبکه عصبی بازگشتی با واحد پنهان GRU

شکل ۹: لایه‌های تشکیل‌دهنده سه معماری بازگشتی عمیق در مدل پیشنهادی

لایه‌های تعریف شده در سه معماری فوق به شرح زیر است:

۱) لایه ورودی در هر سه شبکه وظیفه دریافت ورودی را بر عهده دارد. این لایه، داده‌ها را به صورت یک‌بعدی و توالی دریافت می‌کند. در سه معماری بازگشتی بسیار عمیق پیشنهاد شده است. در این روش، لایه ورودی، داده‌های ورودی را به روش "Zero Center" نرمال‌سازی می‌کند. نرمال‌سازی به روش zero center در رابطه ۵ بیان شده است.

$$Zerocenter(x) = \frac{x - \mu(x)}{\sigma\,(x)} \qquad (۵)$$

در رابطه بالا نماد σ بیانگر انحراف از معیار است و $\mu(x)$ بیانگر میانگین داده ورودی است. لایه ورودی، ورودی‌ها را به صورت توالی نرمال شده به سه لایه LSTM، GRU و یا BiLSTM ارائه می‌دهد.

۲) لایه پنهان که در سه گراف معماری‌های پیشنهادی با سه اسم واحد پنهان LSTM، GRU، BiLSTM نشان داده شده، وابستگی‌های طولانی در داده‌های توالی ورودی را یاد می‌گیرد. این لایه با توجه به واحد پنهان استفاده شده در آن، نام متفاوتی در معماری‌ها دارد. در شبکه‌های پیشنهادی، تعداد واحد پنهان برای هر سه لایه واحد پنهان درج شده در هر سه شبکه برابر با ۲۵۶ در نظر گرفته شده است. عملکرد این لایه‌ها، بر همگرایی شبکه‌ها و بهبود جریان گرادیان توالی‌های طولانی بسیار مؤثر است. در هر یک از لایه‌های پنهان درج شده در شبکه‌های این پژوهش، تابع فعال‌ساز تانژانت هایپربولیک[1] تعریف شده است.

۳) لایه حذف تصادفی در گراف شبکه‌ها پس از هر لایه با واحد پنهان درج شده است. درج لایه‌های حذف تصادفی پس از هر لایه با واحد پنهان باعث تشکیل شبکه‌های بسیار عمیق می‌شود. این لایه‌ها با یک نرخ مشخص که در شبکه‌های پیشنهادی برابر با ۰/۵ تعیین شده است، برای منظم کردن یادگیری شبکه بازگشتی، تعدادی از واحدهای پنهان را به صورت تصادفی حذف می‌کند. حذف تصادفی واحدهای پنهان باعث می‌شود، تعمیم‌پذیری شبکه افزایش پیدا کند و مشکل بیش برازش[2] نیز مرتفع شود. حذف برخی از واحدهای پنهان به صورت تصادفی، چندین شبکه کوچک از شبکه

1. Tanh
2. Overfitting

کلی، تولید می‌کند. در نهایت یکی از شبکه‌های کوچک که دارای وزن کمتری در مقایسه با دیگر شبکه‌ها است به عنوان مناسب‌ترین نماینده انتخاب می‌شود.

۴) لایه تمام متصل به عنوان ورودی تعداد کلاس را دریافت می‌کند که در این مدل برابر با ۲ است. لایه‌های تمام متصل عملکردی مشابه با شبکه‌های عصبی پیشخور دارند. در این لایه، ورودی‌های دریافت شده با وزن‌هایی ضرب می‌شود و با بردار مقادیر ثابت بایاس جمع می‌شود.

۵) لایه SoftMax با استفاده از رابطه ۶ نرخ عضویت هر داده در هر کلاس را محاسبه می‌کند.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_i}} \qquad (۶)$$

در این رابطه، z بیانگر داده ورودی است و k ابعاد آن را نشان می‌دهد.

۶) آخرین لایه در هر سه شبکه، لایه طبقه‌بندی است که وظیفه تخصیص داده‌ها به کلاسی با بیشترین احتمال عضویت را دارد. برای آموزش سه شبکه سایز داده‌ها[1] برابر با ۱۲۸ در نظر گرفته شد. تعداد گردش شبکه[2] در مرحله یادگیری ۳۵۰ تعریف شد. همچنین تعداد واحد پنهان برای هر لایه ۲۵۶ و نرخ حذف تصادفی ۰/۵ در نظر گرفته شد. از داده‌های اعتبارسنجی[3] برای ارزیابی شبکه در حین فرآیند یادگیری استفاده شد. داده‌ها با نرخ ۹۰٪ و ۱۰٪ برای آموزش و ارزیابی شبکه‌ها تقسیم شدند.

تجزیه‌وتحلیل یافته‌ها

برای ارزیابی مدل، از پنج معیار ارزیابی روش‌های طبقه‌بندی شامل صحت، دقت، فراخوانی یا حساسیت، تشخیص و میانگین هارمونیک F استفاده شده که روابط آن‌ها در زیر بیان شده است:

$$\text{صحت} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (۱)$$

$$\text{دقت} = \frac{TP}{TP + FP} \qquad (۲)$$

$$\text{فراخوانی} = \frac{TP}{TP + FN} \qquad (۳)$$

$$\text{میانگین هارمونیک} = \frac{2 * precision * Recall}{precision + Recall} \qquad (۴)$$

$$\text{تشخیص} = \frac{TN}{TN + FP} \qquad (۵)$$

از پایگاه داده Pima که پایگاه دیابت بیماران هندی است برای ارزیابی روش‌ها استفاده می‌شود. این داده از مخزن UCI قابل دریافت است و به صورت آزاد در دسترس عموم است. این پایگاه شامل ۸ ویژگی و یک ویژگی کلاس (ویژگی هدف: دیابت، سالم) و ۷۶۸ نمونه داده است که از این بین، ۲۶۸ داده متعلق به کلاس دیابت و ۵۰۰ داده متعلق به کلاس سالم است. این تعداد به خوبی بیانگر نامتوازن بودن این پایگاه داده است. این پایگاه داده دو چالش اساسی دارد:۱) دارای داده‌های از دست رفته یا گمشده است. ۲) داده‌های کلاس سالم و کلاس بیمار متوازن نیستند.

بررسی تأثیر نوع واحد پنهان

نوع واحد پنهان به دلیل معماری درونی متفاوت آن در نتایج مدل می‌تواند مؤثر واقع شود. نتایج این آزمایش برای هر واحد در ۱۰ اجرا و در جدول ۲ گزارش می‌شود. نتایج این سه جدول نشان می‌دهد که تقریباً در مسئله ما، نتایج هر سه واحد پنهان نزدیک است. در بین سه نوع واحد پنهان بهترین صحت در میانگین ۱۰ اجرای مختلف توسط واحد پنهان LSTM حاصل شده که برابر با ۹۱/۲۱٪ است. بهترین دقت توسط واحد پنهان GRU حاصل شده است که برابر با

1. Mini batch size
2. Epoch
3. Validation

۹۳/۷۴٪ است و بهترین فراخوانی نیز توسط واحد پنهان LSTM به دست آمده که برابر با ۸۸/۰۹٪ است. بهترین تشخیص و میانگین هارمونیک هم توسط BiLSTM حاصل شده است که به ترتیب برابر با ۹۴/۱۶٪ و ۹۰/۴۳٪ است.

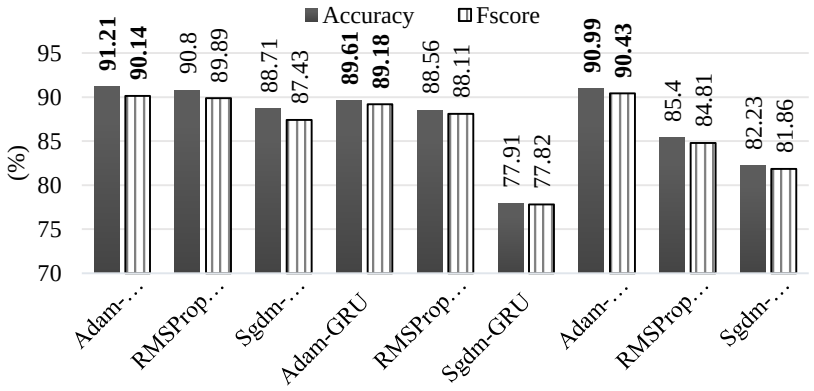
جدول ۲: بررسی نوع واحد پنهان

	واحد پنهان LSTM					واحد پنهان BiLSTM					واحد پنهان GRU				
	صحت	دقت	فراخوانی	میانگین هارمونیک	تشخیص	صحت	دقت	فراخوانی	میانگین هارمونیک	تشخیص	صحت	دقت	فراخوانی	میانگین هارمونیک	تشخیص
۱	۹۲/۹	۹۸	۸۵/۷	۹۱/۴۴	۹۸/۶	۹۰	۹۱/۲	۸۶/۷	۸۸/۸۹	۹۲/۹	۸۸/۵	۹۶/۱	۷۹	۸۶/۷۲	۹۷/۱
۲	۹۱/۵	۹۱/۸	۹۰/۳	۹۱/۰۴	۹۲/۶	۹۲/۳	۹۸/۴	۸۷/۳	۹۲/۵۲	۹۸/۳	۹۰/۷۷	۹۲/۵	۸۹/۹	۹۱/۱۸	۹۱/۸
۳	۹۰	۸۷/۷	۹۱/۹	۸۹/۷۵	۸۸/۲	۸۹/۲	۹۴/۱	۸۱/۴	۸۷/۲۹	۹۵/۸	۸۶/۹	۹۲/۱	۸۲/۹	۸۷/۲۶	۹۱/۷
۴	۹۲/۳	۹۸/۳	۸۴/۴	۹۱/۹۷	۹۸/۴	۹۱/۵	۹۱/۴	۸۹/۸	۹۰/۵۹	۹۳	۹۰	۹۰/۲	۸۸/۷	۸۹/۴۴	۹۱/۲
۵	۹۱/۵	۹۲/۷	۸۷/۹	۹۰/۲۴	۹۴/۴	۸۹/۲	۸۸/۲	۹۰/۹	۸۹/۵۳	۸۷/۵	۹۳/۱	۹۵/۹	۸۷	۹۱/۲۳	۹۷/۴
۶	۹۴/۶	۹۶/۵	۹۱/۷	۹۴/۰۴	۹۷/۱	۹۱/۵	۱۰۰	۸۵/۱	۹۱/۹۵	۱۰۰	۹۱/۵	۹۸/۵	۸۶/۵	۹۲/۱۱	۹۸/۲
۷	۹۳/۱	۹۳/۴	۹۱/۹	۹۲/۶۴	۹۴/۱	۹۳/۱	۸۹/۷	۹۶/۸	۹۳/۱۱	۸۹/۶	۸۴/۶	۸۵/۷	۸۳/۱	۸۴/۳۸	۸۶/۲
۸	۸۶/۲	۸۵/۵	۸۲/۵	۸۳/۹۷	۸۹	۹۱/۵	۹۲/۷	۸۷/۹	۹۰/۲۴	۹۴/۹	۹۳/۱	۹۸/۴	۸۸/۴	۹۳/۱۳	۹۸/۴
۹	۹۱/۵	۸۹/۸	۸۹/۸	۸۹/۸۰	۹۲/۷	۹۰/۸	۹۳	۸۶/۹	۸۹/۸۵	۹۴/۲	۸۸/۵	۹۴/۲	۸۰/۳	۸۶/۷۰	۹۵/۷
۱۰	۸۸/۵	۹۰/۶	۸۲/۸	۸۶/۵۲	۹۳/۱	۹۰/۸	۹۴/۹	۸۶/۲	۹۰/۳۴	۹۵/۴	۸۹/۲	۹۳/۸	۸۵/۹	۸۹/۶۸	۹۳/۲
میانگین	۹۱/۲۱	۹۲/۴۳	۸۸/۰۹	۹۰/۱۴	۹۳/۸۲	۹۰/۹۹	۹۳/۳۶	۸۷/۹۰	۹۰/۴۳	۹۴/۱۶	۸۹/۶۲	۹۳/۷۴	۸۵/۱۷	۸۹/۱۸	۹۴/۰۹

نتایج فوق نشان می‌دهد که تقریباً هر سه واحد پنهان در معیاری منجر به موفقیت مدل در تشخیص دیابت شده است و در نتیجه هر سه واحد پنهان در داده‌های ما دارای عملکرد تقریباً نزدیکی هستند. با این حال با توجه به اینکه واحد پنهان LSTM به صحت بالاتری در مقایسه با دیگر واحدهای پنهان دست یافته است، می‌توان گفت این واحد پنهان در داده‌های دیگر نیز می‌تواند عملکرد قابل اطمینان‌تری ارائه دهد.

بررسی تابع بهینه‌سازی شبکه

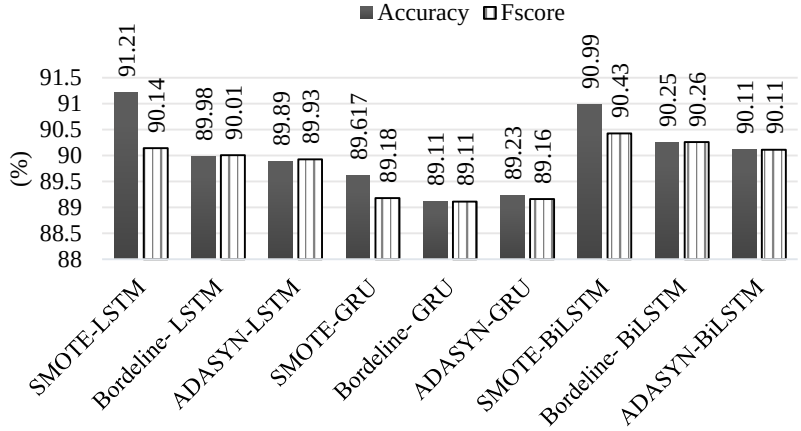
سه تابع بهینه‌سازی Adam, Sgdm و RmsProp برای آموزش شبکه‌های عمیق وجود دارد که هر سه با سه نوع واحد پنهان آزمایش شده‌اند و نتایج آن در شکل ۱۰ گزارش شده است. نتایج این آزمایش نیز به خوبی نشان می‌دهد که در بین سه نوع تابع بهینه‌ساز تابع Adam در مقایسه با دو تابع دیگر عملکرد بهتری در تشخیص دیابت داشته است. در آزمون شبکه LSTM با تابع Adam، توانسته است به صحت ۹۱/۲۱٪ در میانگین ۱۰ اجرا دست پیدا کند. ترکیب Adam با دو واحد GRU و BiLSTM هم به ترتیب باعث حصول صحت ۸۹/۶۱٪ و ۹۰/۹۹٪ شده که در مقایسه با توابع دیگر بالاتر است.



شکل ۱۰: بررسی تأثیر سه تابع بهینه‌ساز شبکه عمیق

بررسی نوع الگوریتم بیش نمونه‌گیری

در این مدل از سه الگوریتم SMOTE و دو نسخه بهبود یافته آن یعنی Borderline SMOTE و ADASYN برای بیش نمونه‌گیری کلاس اقلیت استفاده شده است. هدف از انجام این آزمایش بررسی این موضوع است که کدام نوع الگوریتم بیش نمونه‌گیری منجر به موفقیت بیشتر مدل می‌شود. نتایج این آزمایش در شکل ۱۱ گزارش شده است که نشان می‌دهد، در بین سه الگوریتم بیش نمونه‌گیری، الگوریتم SMOTE داده‌های مصنوعی دقیق‌تری در مقایسه با دو روش دیگر تولید کرده است. هرچه داده‌های مصنوعی با دقت بهتری تولید شوند، باعث می‌شود شباهت بیشتری به داده‌های کلاس اقلیت داشته باشند، در نتیجه شبکه عصبی عمیق با دقت بیشتری می‌تواند آن‌ها را طبقه‌بندی کند. با توجه به کلیه آزمون‌های انجام شده تا این بخش مشخص شد که هر سه نوع واحد پنهان می‌توانند منجر به حصول موفقیت در بخشی از مدل و در معیار مشخصی شوند و از طرف دیگر مشخص شد بهترین تابع بهینه‌ساز Adam و بهترین الگوریتم بیش نمونه‌گیری SMOTE است.



شکل ۱۱: مقایسه نتایج سه الگوریتم بیش نمونه‌گیری

مقایسه نتایج روش‌ها

در جدول ۳ نتایج مقایسه بین مدل‌های این پژوهش و دو مدل رامش و چنگ به ترتیب در سال‌های ۲۰۲۱ و ۲۰۲۲ انجام شده است.

مقایسه نتایج نشان می‌دهد که شبکه عصبی بازگشتی با هر سه نوع واحد پنهان در مقایسه با روش‌های یادگیری ماشین سنتی استفاده شده در دو روش رامش و چنگ به موفقیت بیشتری رسیده‌اند. استفاده از معماری عمیق برای تشخیص

دیابت یکی از عوامل موفقیت مدل پیشنهادی است. در واقع نتیجه مطالعات مختلف نشان می‌دهد که در اکثر حوزه‌های پزشکی، یادگیری عمیق در مقایسه با یادگیری ماشین به موفقیت بیشتری دست پیدا می‌کند. نتایج این پژوهش نیز در راستای تحقیقات دیگر نشان می‌دهد که در حوزه تشخیص دیابت نیز روش‌های یادگیری عمیق در مقایسه با روش‌های یادگیری ماشین سنتی عملکرد بهتری دارند.

یکی دیگر از دلایل موفقیت مدل پیشنهادی در مقایسه با روش‌های دیگر مرحله پیش‌پردازش است. پیش‌پردازش داده‌ها یکی از عواملی است که به خوبی می‌تواند منجر به موفقیت مدل شود. در مدل پیشنهادی حذف داده‌های پرت و همچنین مقداردهی به مقادیر گمشده دو مرحله پیش‌پردازشی بودند که در موفقیت مدل در مقایسه با روش‌های دیگر مؤثر بودند. از دیگر عوامل موفقیت مدل، مرحله بیش نمونه‌گیری است که در آن از افزایش داده‌ها بر روی کل داده‌ها استفاده شده، در حالی که در روش رامش از افزایش داده فقط بر روی داده‌های آموزش استفاده شده است. نتایج حاصل نشان می‌دهد که این مدل نتوانسته در تشخیص دیابت چندان دقیق عمل کند؛ به این دلیل که مرحله پیش‌پردازش مناسبی ندارد و از روش‌های یادگیری ماشین سنتی استفاده کرده است؛ اما در مقابل در مدل پیشنهادی این پژوهش مرحله پیش‌پردازش و بیش نمونه‌گیری و شبکه عصبی عمیق با یکدیگر ترکیب شده‌اند که بهترین نتایج را در تشخیص دیابت ارائه دهند.

جدول ۳: مقایسه نتایج مدل‌ها در Pima

	صحت	دقت	فراخوانی	میانگین هارمونیک	تشخیص
Propose LSTM	**۹۱/۲۱**	۹۲/۴۳	۸۸/۰۹	۹۰/۱۴	۹۳/۸۲
Propose GRU	۸۹/۶۱	**۹۳,۷۴**	۸۵/۱۷	۸۹/۱۸	۹۴/۰۹
Propose BiLSTM	۹۰/۹۹	۹۳/۳۶	۸۷/۹	**۹۰/۴۳**	**۹۴/۱۶**
Ramesh et al. 2021 (KNN)	۷۹/۸	۸۲/۱	۸۷/۲	۸۴/۵۷	۷۲/۴
Ramesh et al.2021 (SVM-RBF)	۸۳/۲	۸۵/۵	۸۷/۳	۸۶/۳۹	۷۹
Chang et al. 2022 (J48)	۷۵/۶۵	۷۰/۸۶	**۸۹/۹۲**	۷۹/۲۶	۶۰/۳۶
Chang et al. 2022 (Random Forest)	۷۳/۹۱	۸۰/۷۹	۷۹/۷۴	۸۰/۲۶	۶۲/۳۴
Chang et al. 2022 (NB)	۷۷/۸۳	۸۱/۲۵	۸۶/۰۹	۸۳/۶۰	۶۲/۰۳

نتایج فوق نشان می‌دهد که شبکه عصبی عمیق پیشنهادی با واحد پنهان LSTM در مقایسه با روش رامش صحت و دقت را در مقایسه با k نزدیک‌ترین همسایه به ترتیب ۱۱/۴۱٪ و ۱۰/۳۳٪ بهبود بخشیده است. همچنین این شبکه در مقایسه با ماشین بردار پشتیبان صحت و دقت را به ترتیب ۸/۰۱٪ و ۶/۹۳٪ افزایش داده است. در معیار تشخیص نیز شبکه LSTM در مقایسه با ماشین بردار پشتیبان و k نزدیک‌ترین همسایه بالغ بر ۱۴٪ موفق‌تر بوده است. شبکه GRU در مقایسه با k نزدیک‌ترین همسایه صحت را ۹/۸۲٪ و دقت را ۱۱/۶۴٪ بهبود داده است. شبکه بازگشتی GRU در مقایسه با ماشین بردار پشتیبان صحت و دقت را بیش از ۶٪ افزایش داده است. شبکه بازگشتی BiLSTM نیز در مقایسه با k نزدیک‌ترین همسایه صحت و دقت را بیش از ۱۱٪ و در مقایسه با ماشین بردار پشتیبان بیش از ۷٪ دقت و صحت تشخیص دیابت را بهبود بخشیده است. موفقیت شبکه‌های عصبی در مقایسه با روش چنگ بیشتر است و در مقایسه با آن بالغ بر ۱۳٪ در بیشتر معیارهای ارزیابی موفقیت حاصل شده است. روش چنگ انتخاب ویژگی را به درستی انجام

نداده است و ۳ ویژگی منتخب آن دانش کافی از اطلاعات بیماران ارائه نمی‌دهد. نتایج فوق به وضوح نشان می‌دهد که مدل به طور میانگین در مقایسه با ۵ الگوریتم یادگیری ماشین سنتی توانسته است تشخیص دیابت را با موفقیت بیشتر انجام دهد. خلاصه نتایج نشان می‌دهد بین ۶٪ تا ۱۷٪ سه شبکه عصبی بازگشتی با سه واحد پنهان توانسته‌اند صحت تشخیص دیابت را افزایش دهند. همچنین دقت سه شبکه عصبی بازگشتی با سه واحد پنهان مختلف در مقایسه با ۵ روش دیگر بین ۶٪ تا ۲۲٪ افزایش داشته است.

نتیجه‌گیری و کارهای آتی

پژوهش حاضر یک سیستم دستیار پزشک در حوزه تشخیص دیابت معرفی کرده است. دیابت از جمله بیماری‌هایی است که عدم کنترل آن به کلیه اندام‌های داخلی بدن آسیب جبران‌ناپذیری را وارد می‌کند. از این رو تشخیص به موقع آن در راستای کنترل و انتخاب یک روش درمانی مناسب می‌تواند در افزایش بقای بیماران و افزایش کیفیت زندگی آن‌ها مؤثر واقع شود. تاکنون روش‌های مختلفی مبتنی بر روش‌های یادگیری ماشین سنتی در این زمینه ارائه شدند؛ اما اغلب آن‌ها به مشکل عدم توازن داده‌ها بی‌توجه بودند، در نتیجه نتایج امیدوار کننده‌ای ارائه ندادند. در این پژوهش یک روش تشخیص دیابت چندمرحله‌ای ارائه شد که دارای چندین مرحله پیش‌پردازش مثل حذف داده‌های پرت و مقداردهی به مقادیر گمشده است. همچنین در آن مشکل عدم توازن داده‌ها با استفاده از بیش نمونه‌گیری مرتفع شده است و با استفاده از سه شبکه عصبی بازگشتی تشخیص دیابت انجام شده است. نتایج آزمایش‌های مختلف در داده‌های دیابت هند Pima نشان داد که سه واحد پنهان مختلف در داده‌های دیابت دارای عملکرد تقریباً مشابهی هستند. همچنین برای یادگیری شبکه عصبی تابع Adam دارای بهترین نتایج است. مشخص شد که در بین روش‌های بیش نمونه‌گیری الگوریتم SMOTE منجر به حصول نتایج بهتری در مقایسه با دو نسخه بهبود یافته آن می‌شود. مقایسه نتایج با ۵ روش یادگیری ماشین سنتی نشان داد که روش‌های یادگیری عمیق در مقایسه با روش‌های یادگیری ماشین سنتی مثل جنگل تصادفی، ماشین بردار پشتیبان، k نزدیک‌ترین همسایه، بیز ساده و درخت تصمیم J48 در تشخیص دیابت به طور چشمگیری می‌تواند موفق‌تر عمل کند. روش یادگیری عمیق در داده‌های کم هم نتایج قابل قبولی را ارائه می‌دهد. بالاخص واحد پنهان GRU به صورت تخصصی در شبکه‌های بازگشتی بر روی داده‌های کم می‌تواند موفق عمل کند. در ادامه این پژوهش برای بهبود نتایج، می‌توان یک مرحله انتخاب ویژگی مبتنی بر روش‌های فرااکتشافی مثل الگوریتم گرگ خاکستری به مدل اضافه کرد. استفاده از چند شبکه‌عصبی مختلف مثل شبکه چندلایه پرسپترون و شبکه‌های بازگشتی به صورت ترکیبی برای تشخیص دیابت نیز می‌تواند بررسی شود. همچنین می‌توان سه شبکه عصبی بازگشتی را با استفاده از رأی‌گیری اکثریت ترکیب و به صورت ترکیبی تشخیص دیابت را انجام داد.

منابع

Abedini, M, Bijari, A, & Banirostam, T. (2020). Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network.

Chang, V, Bailey, J, Xu, Q. A, & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, 1-17.

Chawla, N.V, Bowyer, K. W, Hall, L. O, & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Choubey, D. K, Kumar, M, Shukla, V, Tripathi, S, & Dhandhania, V. K. (2020). Comparative Analysis of Classification Methods with PCA and LDA for Diabetes. *Current Diabetes Reviews*, *16*(8), 833-850.

Choudhury, A, & Gupta, D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. In *Recent Developments in Machine Learning and Data Analytics* (pp. 67-78). Springer.

Cui, Z, Ke, R, Pu, Z, & Wang, Y. (2018). Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. *arXiv preprint*, 1-11.

Das, H., Naik, B, & Behera, H. (2018). Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. In *Progress in computing, analytics and networking* (pp. 539-549). Springer.

Diabetes, U. (2015). Number of people with diabetes up 60 per cent in last decade. In.

Fernández, A, Garcia, S, Herrera, F, & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863-905.

Fiarni, C, Sipayung, E. M, & Maemunah, S. (2019). Analysis and prediction of diabetes complication disease using data mining algorithm. *Procedia Computer Science*, *161*, 449-457.

Ganesh, P. S, & Sripriya, P. (2019). A Comparative Review of Prediction Methods for Pima Indians Diabetes Dataset. International Conference On Computational Vision and Bio Inspired Computing,

Garga, S. B, Mahajanb, A. K, & Kamalc, T. (2017). An Approach for Diabetes Detection Using Data Mining Classification Techniques. *International Journal of Engineering Sciences*, 202-218.

Gers, F. A, Schmidhuber, J, & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation*, *12*(10), 2451-2471.

Ghorbani, R, & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. *International Journal of Data Network Science*, *3*(2), 47-70.

Haibo, H, Yang, B, Garcia, E. A, & Shutao, L. (2008, 1-8 June 2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence),

Hairani, H, Saputro, K. E, & Fadli, S. (2020). K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4. 5, SVM, dan naive Bayes. *Jurnal Teknologi dan Sistem Komputer*, *8*(2), 89-93.

Han, H, Wang, W.-Y, & Mao, B.-H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International conference on intelligent computing,

Hochreiter, S, & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Kandhasamy, J. P, & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, *47*, 45-51.

Kannadasan, K, Edla, D. R, & Kuppili, V. (2019). Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clinical Epidemiology Global Health*, *7*(4), 530-535.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232.

kumar Dewangan, A, & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. *International Journal of Engineering Applied Sciences*, *2*(5), 145-148.

López, F. (2021). *SMOTE: Synthetic Data Augmentation for Tabular Data*. https://towardsdatascience.com/smote-synthetic-data-augmentation-for-tabular-data-1ce28090debc

Meng, X.-H, Huang, Y.-X, Rao, D.-P, Zhang, Q, & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, *29*(2), 93-99.

Naz, H. & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes Metabolic Disorders*, *19*(1), 391-403.

Nilashi, M, bin Ibrahim, O, Ahmadi, H, & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers Chemical Engineering*, *106*, 212-223.

Nnamoko, N, & Korkontzelos, I. (2020). Efficient treatment of outliers and class imbalance for diabetes prediction. *Artificial Intelligence in Medicine*, *104*, 1-12.

Organization, W. H. (2019). Classification of diabetes mellitus. 1-40.

Pertiwi, A, Bachtiar, N, Kusumaningrum, R, Waspada, I, & Wibowo, A. (2020). Comparison of performance of k-nearest neighbor algorithm using smote and k-nearest neighbor algorithm without smote in diagnosis of diabetes disease in balanced data. Journal of Physics: Conference Series,

Perveen, S, Shahbaz, M, Guergachi, A, & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, *82*, 115-121.

Rajni, R, & Amandeep, A. (2019). RB-Bayes algorithm for the prediction of diabetic in Pima Indian dataset. *International Journal of Electrical Computer Engineering*, *9*(6), 4866-4872.

Ramesh, J, Aburukba, R, & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters*, *8*(3), 45-57.

Roglic, G. (2016). WHO Global report on diabetes: A summary [Review Article]. *1*(1), 3-8. https://www.ijncd.org/article.asp?issn=2468-8827;year=2016;volume=1;issue=1;spage=3;epage=8;aulast=Roglic

Santhanam, T, & Padmavathi, M. (2015). Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. *Procedia Computer Science*, *47*, 76-83.

Saxena, K, Khan, Z, & Singh, S. (2014). Diagnosis of diabetes mellitus using k nearest neighbor algorithm. *International Journal of Computer Science Trends Technology*, *2*(4), 36-43.

Shuja, M, Mittal, S, & Zaman, M. (2020). Effective prediction of type ii diabetes mellitus using data mining classifiers and SMOTE. In *Advances in Computing and Intelligent Systems* (pp. 195-211). Springer.

Sisodia, D, & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science*, *132*, 1578-1585.

Sreejith, S, Nehemiah, H. K, & Kannan, A. (2020). Clinical Data Classification Using an Enhanced SMOTE and Chaotic Evolutionary Feature Selection. *Computers in Biology Medicine*, *126*, 1-14.

Wang, Q, Cao, W, Guo, J, Ren, J, Cheng, Y, & Davis, D. (2019). DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values. *IEEE Access*, *7*, 102232-102238.

Wu, H, Yang, S, Huang, Z, He, J, & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100-107.

Xu, W, Zhang, J, Zhang, Q, & Wei, X. (2017). Risk prediction of type II diabetes based on random forest model. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB),

Yu, Y, Si, X, Hu, C, & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, *31*(7), 1235-1270.

Zhai, J, Zhang, S, & Wang, C. (2017). The classification of imbalanced large data sets based on mapreduce and ensemble of elm classifiers. *International Journal of Machine Learning Cybernetics*, *8*(3), 1009-1017.