پردازش اطلاعات کوانتومی

نام و نام خانوادگی: بهزاد جنتی

۸۱۰۱۰۳۰۹۸

گزارش پروژه نهایی

چکیده

در این پروژه، یک شبکه عصبی کوانتومی (QNN) بر اساس معماری Farhi-Neven پیاده‌سازی شده است که هدف آن پیش‌بینی تأثیر جهش‌های ژنتیکی بر اتصال داروها به پروتئین‌های هدف می‌باشد. این کار بر اساس مقاله HypaCADD انجام شده که یک جریان کاری ترکیبی کلاسیک-کوانتومی برای طراحی دارو ارائه می‌دهد. مدل پیشنهادی با استفاده از کتابخانه Qiskit پیاده‌سازی شده و روی مجموعه داده GenoDock آموزش داده شده است. نتایج نشان می‌دهد که با استفاده از بهینه‌ساز SPSA و ۲۰۰ تکرار آموزش، دقت ۹۱ درصد روی مجموعه داده اعتبارسنجی حاصل شده است که نشان‌دهنده عملکرد قابل قبول شبکه‌های عصبی کوانتومی در مسائل طبقه‌بندی زیست‌پزشکی است.

کلمات کلیدی: یادگیری ماشین کوانتومی، شبکه عصبی کوانتومی، طراحی دارو، پیش‌بینی جهش، Qiskit

۱ مقدمه

۱.۱ پیش‌زمینه و انگیزه

محاسبات کوانتومی به عنوان یک پارادایم محاسباتی نوین، پتانسیل بالایی برای حل مسائل پیچیده‌ای دارد که برای کامپیوترهای کلاسیک دشوار یا غیرممکن هستند. یکی از حوزه‌های کاربردی مهم محاسبات کوانتومی، طراحی دارو به کمک کامپیوتر (CADD:Computer-Aided Drug Design) است. در این حوزه، مسائل ترکیبیاتی و چندجسمی وجود دارند که با افزایش اندازه سیستم، پیچیدگی محاسباتی آن‌ها به صورت نمایی رشد می‌کند.

بیت‌های کوانتومی (کیوبیت‌ها) فضای حالتی نمایی بزرگ‌تر از بیت‌های کلاسیک فراهم می‌کنند. در حالی که یک بیت کلاسیک تنها دو حالت گسسته ۰ و ۱ دارد، یک کیوبیت می‌تواند در برهم‌نهی این دو حالت قرار گیرد. این ویژگی به همراه درهم‌تنیدگی کوانتومی، امکان پردازش موازی گسترده‌ای را فراهم می‌کند.

۲.۱ معرفی مقاله مرجع

این پروژه بر اساس مقاله «Insights from incorporating quantum computing into drug design workflows» که در سال ۲۰۲۳ در مجله Bioinformatics منتشر شده، انجام گرفته است . این مقاله یک جریان کاری ترکیبی کلاسیک-کوانتومی به نام HypaCADD را برای طراحی دارو ارائه می‌دهد.
هدف اصلی HypaCADD یافتن لیگاندهایی (مولکول‌های دارویی) است که:

۱. به پروتئین هدف با میل ترکیبی بالا متصل شوند

۲. در برابر جهش‌های ژنتیکی پروتئین مقاوم باشند

مطالعه موردی این مقاله روی پروتئین CLpro۳ ویروس SARS-CoV-2 انجام شده است. این پروتئین نقش کلیدی در تکثیر ویروس دارد و هدف مناسبی برای داروهای ضدویروسی محسوب می‌شود.

۳.۱ رویکرد ترکیبی کلاسیک-کوانتومی

جریان کاری HypaCADD از چهار مرحله اصلی تشکیل شده است:

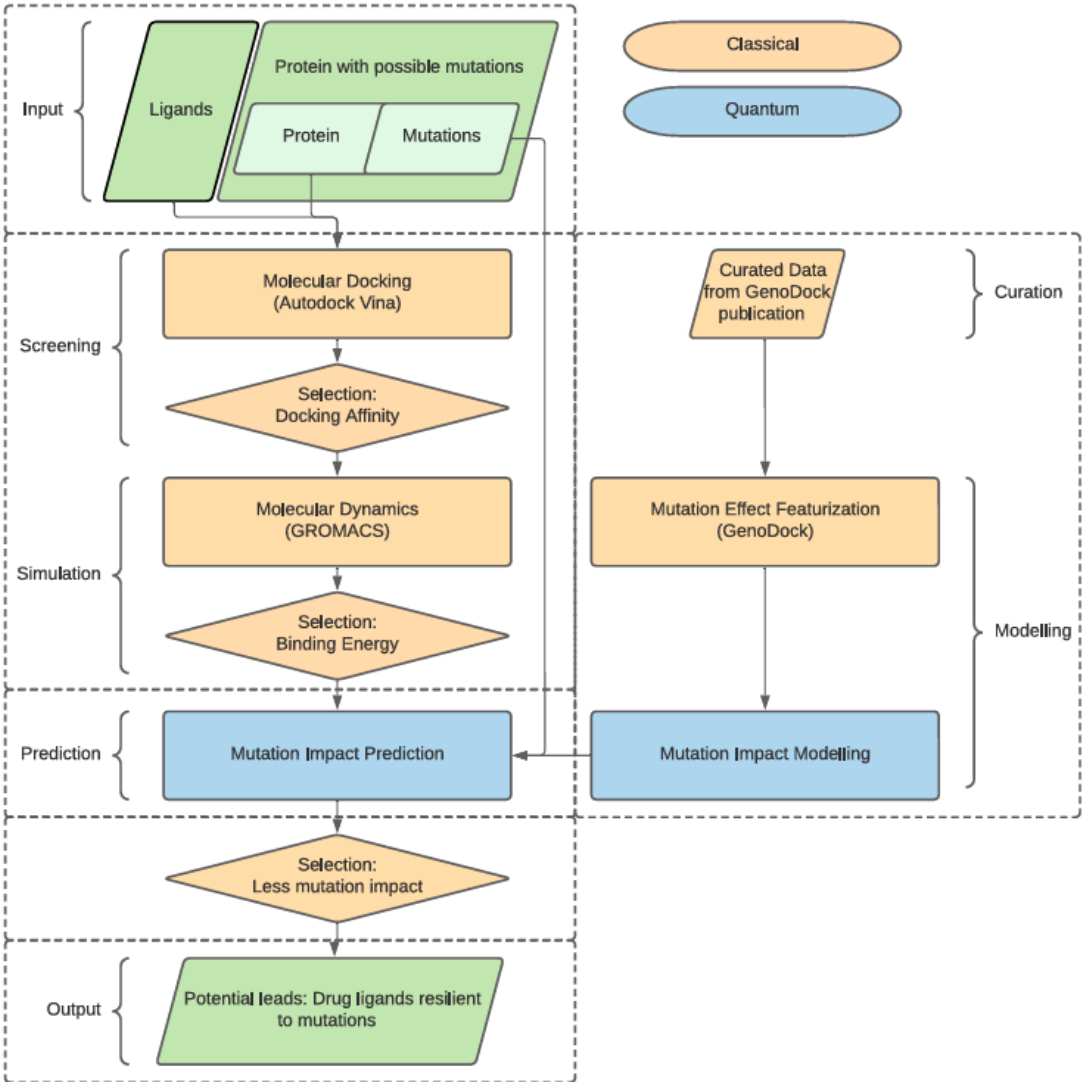
۱. **Molecular Docking** (کلاسیک): غربالگری اولیه داروهای کاندید با استفاده از AutoDock Vina

۲. **Molecular Dynamics** (کلاسیک): شبیه‌سازی دقیق‌تر اتصال دارو-پروتئین با GROMACS

۳. **Feature Extraction** (کلاسیک): استخراج ویژگی‌ها از هر جفت جهش-دارو

۴. **QML Prediction** (کوانتومی): پیش‌بینی تأثیر جهش با شبکه عصبی کوانتومی

نویسندگان مقاله به صورت غیرشهودی تشخیص داده‌اند که ماژول پیش‌بینی تأثیر جهش، بهترین کاندید برای جایگزینی با محاسبات کوانتومی است. دلیل این انتخاب، مستقل بودن این ماژول و قابلیت مقایسه مستقیم با روش‌های کلاسیک است.



شکل ۱: جریان کاری HypaCADD: ترکیب محاسبات کلاسیک و کوانتومی برای طراحی دارو. مستطیل‌های آبی نشان‌دهنده بخش‌های کوانتومی هستند.

۲ مجموعه داده

۱.۲ منبع داده

مجموعه داده مورد استفاده در این پروژه، داده‌های GenoDock است که توسط نویسندگان مقاله HypaCADD تهیه شده است. این داده‌ها شامل ویژگی‌های استخراج شده از جفت‌های جهش-لیگاند برای پروتئین CLpro۳ ویروس SARS-CoV-2 می‌باشد.

۲.۲ ساختار داده

مجموعه داده شامل دو فایل اصلی است:

- `genodock_cluster3_fix_train.tsv`: داده‌های آموزش (۵۱۴۲ نمونه)
- `genodock_cluster3_fix_val.tsv`: داده‌های اعتبارسنجی (۵۱۳۹ نمونه)

۳.۲ ویژگی‌های مورد استفاده

برای سازگاری با سیستم‌های کوانتومی ۵ کیوبیتی IBM، از ۴ ویژگی غیر وابسته به لیگاند استفاده شده است:

جدول ۱: ویژگی‌های مورد استفاده در مدل QNN

ردیف	نام ویژگی	توضیح
۱	`bind_site`	آیا جهش در محل اتصال دارو قرار دارد؟ (۰ یا ۱)
۲	`distance`	فاصله جهش از دارو (آنگستروم)
۳	`polarity_change_index`	شاخص تغییر قطبیت اسید آمینه
۴	`volume_change_index`	شاخص تغییر حجم اسید آمینه

۴.۲ برچسب‌ها و توزیع کلاس‌ها

برچسب هدف (`target`) دارای دو کلاس است:

- کلاس ۰ (**Non-disruptive**): جهش تأثیری بر اتصال دارو ندارد
- کلاس ۱ (**Disruptive**): جهش اتصال دارو را مختل می‌کند

جدول ۲: توزیع کلاس‌ها در مجموعه داده آموزش

کلاس	تعداد نمونه	درصد
Non-disruptive (0)	۴۸۰۶	٪93.5
Disruptive (1)	۳۳۶	٪6.5
مجموع	۵۱۴۲	۱۰۰٪

همان‌طور که مشاهده می‌شود، داده‌ها به شدت نامتوازن هستند (نسبت ۳:۱.۱۴). این موضوع یکی از چالش‌های اصلی در آموزش مدل است.

۳ معماری مدل و مدار کوانتومی

۱.۳ معماری Farhi-Neven

معماری شبکه عصبی کوانتومی مورد استفاده، بر اساس مدل پیشنهادی Farhi و Neven است. این معماری شامل اجزای زیر می‌باشد:

- کیوبیت‌های ورودی: ۴ کیوبیت (یکی به ازای هر ویژگی)
- کیوبیت خروجی (**Readout**): ۱ کیوبیت
- تعداد کل کیوبیت‌ها: ۵
- تعداد لایه‌ها: ۳ ($\mathrm{RZX} \to \mathrm{RXX} \to \mathrm{RZX}$)
- تعداد پارامترهای قابل آموزش: ۱۲

۲.۳ نگاشت ویژگی (Feature Map)

ویژگی‌های ورودی ابتدا به بازه $[0, 1]$ نرمال‌سازی می‌شوند. سپس هر ویژگی با استفاده از گیت R_X به کیوبیت متناظر نگاشت می‌شود:

$$R_X(\phi_i) = e^{-i\frac{\phi_i}{2}X} \quad \text{where} \quad \phi_i = \pi \times x_i \tag{۱}$$

که در آن x_i مقدار نرمال‌شده ویژگی i-ام است.

۳.۳ لایه‌های متغیر (Variational Layers)

مدار از سه لایه متغیر تشکیل شده است:

۱. لایه اول (**RZX**): گیت‌های $R_{ZX}(\theta)$ بین هر کیوبیت ورودی و کیوبیت خروجی

۲. لایه دوم (**RXX**): گیت‌های $R_{XX}(\theta)$ بین هر کیوبیت ورودی و کیوبیت خروجی

۳. لایه سوم (**RZX**): گیت‌های $R_{ZX}(\theta)$ بین هر کیوبیت ورودی و کیوبیت خروجی

گیت‌های دو کیوبیتی به صورت زیر تعریف می‌شوند:

$$R_{ZX}(\theta) = e^{-i\frac{\theta}{2}(Z\otimes X)} \tag{۲}$$

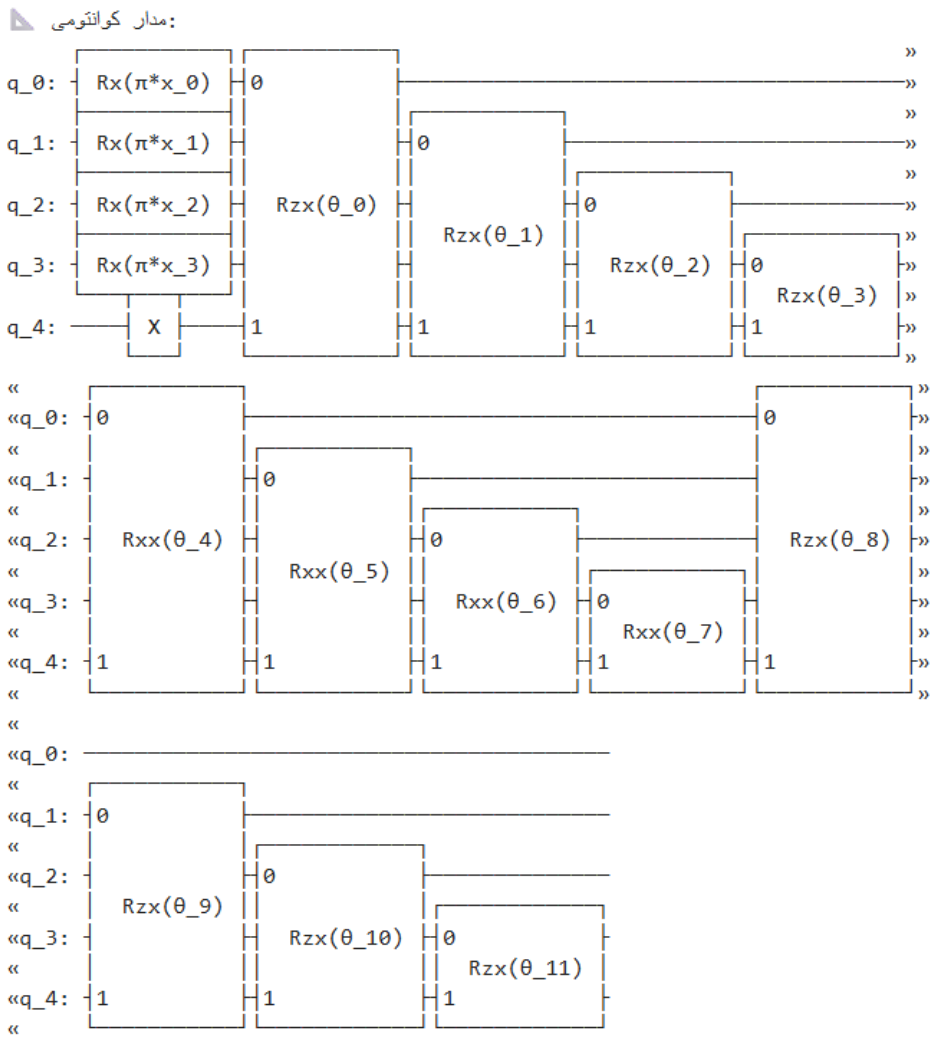
$$R_{XX}(\theta) = e^{-i\frac{\theta}{2}(X\otimes X)} \tag{۳}$$

۴.۳ اندازه‌گیری و خروجی

پس از اجرای تمام لایه‌ها، مؤلفه Y کیوبیت خروجی اندازه‌گیری می‌شود. مقدار انتظاری این اندازه‌گیری به عنوان پیش‌بینی مدل استفاده می‌شود:

$$\text{prediction} = \langle z, 1|U^\dagger(\vec{\theta})Y_{\text{readout}}U(\vec{\theta})|z, 1\rangle \tag{۴}$$

که در آن $|z, 1\rangle$ حالت اولیه سیستم (کیوبیت‌های ورودی بر اساس داده و کیوبیت خروجی در حالت $|1\rangle$) و $U(\vec{\theta})$ حاصل‌ضرب تمام گیت‌های مدار است.



شکل ۲: مدار کوانتومی QNN با ۵ کیوبیت و ۳ لایه. کیوبیت‌های q_0 تا q_3 ورودی و q_4 خروجی است.

۵.۳ تابع هزینه

تابع هزینه مورد استفاده، مطابق با مقاله Farhi-Neven تعریف شده است:

$$\text{loss}(\vec{\theta}, z) = 1 - l(z) \times \text{prediction} \tag{۵}$$

که در آن $l(z)$ برچسب واقعی نمونه z است. این تابع روی تمام نمونه‌های آموزش جمع زده شده و نسبت به پارامترهای $\vec{\theta}$ کمینه می‌شود.

۴ پیاده‌سازی و آموزش

۱.۴ ابزارها و کتابخانه‌ها

پیاده‌سازی این پروژه با استفاده از ابزارهای زیر انجام شده است:

- **Qiskit 0.45.0**: کتابخانه اصلی محاسبات کوانتومی IBM
- **Qiskit Machine Learning 0.6.1**: ماژول یادگیری ماشین کوانتومی
- **Python 3.12**: زبان برنامه‌نویسی
- **NumPy, Pandas, Scikit-learn**: پردازش داده و ارزیابی

۲.۴ پیش‌پردازش داده

مراحل پیش‌پردازش داده شامل موارد زیر است:

۱. بارگذاری فایل‌های TSV

۲. انتخاب ۴ ویژگی غیر وابسته به لیگاند

۳. نرمال‌سازی ویژگی‌ها به بازه $[0, 1]$ با MinMaxScaler

۳.۴ تنظیمات آموزش

دو پیکربندی مختلف برای آموزش مدل استفاده شده است:

جدول ۳: مقایسه دو پیکربندی آموزش

پارامتر	پیکربندی اول	پیکربندی دوم
بهینه‌ساز	COBYLA	SPSA
تعداد تکرار	۲۰۰ (توقف در ۲۰)	۲۰۰
نرخ یادگیری	-	۰/۰۵
Perturbation	-	۰/۰۵
تعداد نمونه آموزش	۵۱۴۲	۵۱۴۲

۴.۴ بهینه‌سازهای مورد استفاده

۱.۴.۴ COBYLA

الگوریتم COBYLA (Constrained Optimization BY Linear Approximation) یک روش بهینه‌سازی بدون گرادیان است که برای مسائل با محدودیت‌های غیرخطی مناسب است. این بهینه‌ساز در پیکربندی اول استفاده شد اما به دلیل وجود callback و عدم بهبود، آموزش در تکرار ۲۰ متوقف شد.

۲.۴.۴ SPSA

الگوریتم SPSA (Simultaneous Perturbation Stochastic Approximation) یک روش بهینه‌سازی تصادفی است که برای شبکه‌های عصبی کوانتومی بسیار مناسب است. این الگوریتم با اختلال همزمان تمام پارامترها، گرادیان را تخمین می‌زند و نسبت به نویز مقاوم است.

۵ نتایج و ارزیابی

۱.۵ نتایج آموزش با COBYLA

در پیکربندی اول با بهینه‌ساز COBYLA، آموزش به دلیل عدم بهبود Loss در تکرار ۲۰ متوقف شد:

جدول ۴: نتایج آموزش با COBYLA

معیار	مقدار
تعداد تکرار	۲۰ (از ۲۰۰)
Loss نهایی	۰/۰۶
دقت آموزش	۵۲/۹۹٪
دقت اعتبارسنجی	۵۲/۳۶٪
زمان آموزش	۵۹۹ ثانیه

۲.۵ نتایج آموزش با SPSA

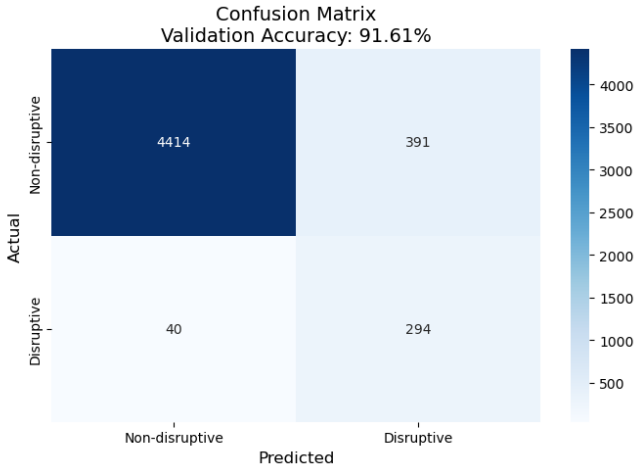
در پیکربندی دوم با بهینه‌ساز SPSA، نتایج به طور قابل توجهی بهبود یافت:

جدول ۵: نتایج آموزش با SPSA

معیار	مقدار
تعداد تکرار	۲۰۰
Loss اولیه	۰/۵۱
Loss نهایی	۰/۱۱
دقت آموزش	۹۱/۵۲٪
دقت اعتبارسنجی	۹۱/۶۱٪
زمان آموزش	۶۸۸۳ ثانیه

۳.۵ روند کاهش Loss

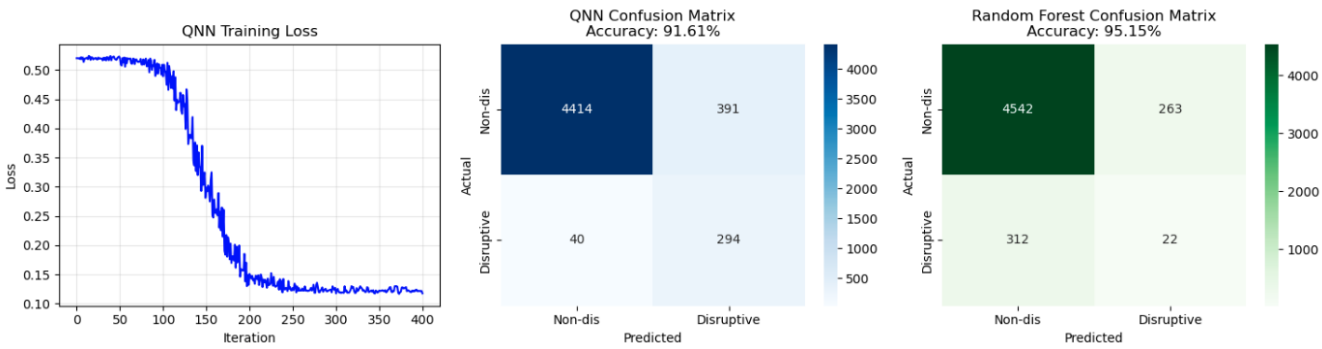
نمودار زیر روند کاهش تابع هزینه در طول آموزش با SPSA را نشان می‌دهد:



شکل ۳: روند کاهش Loss در طول ۲۰۰ تکرار آموزش با SPSA. Loss از ۰/۵۱ به ۰/۱۱ کاهش یافته است.

۴.۵ ماتریس درهم‌ریختگی

ماتریس درهم‌ریختگی نتایج پیش‌بینی روی مجموعه داده اعتبارسنجی:



شکل ۴: ماتریس درهم‌ریختگی نتایج QNN روی داده‌های اعتبارسنجی

۵.۵ گزارش طبقه‌بندی

جدول ۶: گزارش طبقه‌بندی روی داده‌های اعتبارسنجی

کلاس	Precision	Recall	F1-Score	Support
(۰) Non-disruptive	۰/۹۹	۰/۹۲	۰/۹۵	۴۸۰۵
(۱) Disruptive	۰/۴۳	۰/۸۸	۰/۵۸	۳۳۴
Accuracy		۹۱.۰		
Macro Avg	۰/۷۱	۰/۹۰	۰/۷۷	۵۱۳۹
Weighted Avg	۰/۹۵	۰/۹۲	۰/۹۳	۵۱۳۹

۶.۵ مقایسه با نتایج مقاله

جدول ۷: مقایسه نتایج این پروژه با نتایج گزارش شده در مقاله

مدل	این پروژه	مقاله (**GenoDock**)	مقاله (**Platinum**)
QNN (3-layer)	۹۱٪	AUC: ۷۰/۲	AUC: ۶۱/۸
Random Forest	۹۵٪	AUC: ۷۱/۱	AUC: ۶۴/۴
weighted-mQNN	-	AUC: ۷۴/۶۰	AUC: ۶۹/۸

نکته مهم: نتایج گزارش شده در مقاله بر اساس معیار AUC و روی مجموعه داده تست Platinum (داده‌های تجربی واقعی) است، در حالی که نتایج این پروژه بر اساس دقت (Accuracy) و روی مجموعه داده اعتبارسنجی GenoDock گزارش شده است. این تفاوت در معیار ارزیابی و مجموعه داده تست باید در مقایسه‌ها لحاظ شود.

۶ تحلیل و بحث

۱.۶ تأثیر انتخاب بهینه‌ساز

نتایج نشان می‌دهد که انتخاب بهینه‌ساز تأثیر قابل توجهی بر عملکرد مدل دارد:

- **COBYLA**: این بهینه‌ساز در تکرارهای اولیه گیر کرد و نتوانست Loss را به طور مؤثر کاهش دهد. دقت نهایی ٪۵۲ تقریباً معادل حدس تصادفی است.
- **SPSA**: این بهینه‌ساز با نرخ یادگیری ۰۵.۰ و perturbation مناسب، توانست Loss را از ۰/۵۱ به ۰/۱۱ کاهش دهد و دقت ٪۹۱ را حاصل کند.

۲.۶ چالش داده‌های نامتوازن

یکی از چالش‌های اصلی این پروژه، نامتوازن بودن شدید داده‌ها بود (نسبت ۱۴/۳:۱). اگرچه دقت کلی ٪۹۱ به دست آمد، اما Recall کلاس اقلیت (Disruptive) تنها ٪۸۸ است. این نشان می‌دهد که مدل بیشتر به سمت کلاس اکثریت متمایل شده است.

۳.۶ مقایسه با روش‌های کلاسیک

طبق نتایج مقاله، شبکه‌های عصبی کوانتومی عملکردی «هم‌تراز» (on par) با روش‌های کلاسیک مانند Random Forest دارند. این نتیجه مهم است زیرا نشان می‌دهد:

۱. QNN می‌تواند مسائل واقعی زیست‌پزشکی را حل کند

۲. با پیشرفت سخت‌افزار کوانتومی، پتانسیل بهبود وجود دارد

۳. رویکرد ترکیبی کلاسیک-کوانتومی عملی و قابل اجراست

۴.۶ محدودیت‌ها

- اجرا روی شبیه‌ساز کوانتومی (نه سخت‌افزار واقعی)
- عدم استفاده از weighted loss برای داده‌های نامتوازن
- محدودیت در تعداد کیوبیت‌ها (۵ کیوبیت)

۷ نتیجه‌گیری

در این پروژه، یک شبکه عصبی کوانتومی بر اساس معماری Farhi-Neven برای پیش‌بینی تأثیر جهش‌های ژنتیکی بر اتصال دارو پیاده‌سازی شد. نتایج اصلی به شرح زیر است:

۱. پیاده‌سازی موفق معماری QNN با ۵ کیوبیت و ۱۲ پارامتر قابل آموزش
۲. دستیابی به دقت ۹۱٪ روی مجموعه داده اعتبارسنجی با بهینه‌ساز SPSA
۳. تأیید اهمیت انتخاب بهینه‌ساز مناسب برای آموزش QNN
۴. نشان دادن قابلیت QNN در حل مسائل واقعی طبقه‌بندی زیست‌پزشکی

این کار نشان می‌دهد که یادگیری ماشین کوانتومی می‌تواند به عنوان بخشی از جریان‌های کاری طراحی دارو مورد استفاده قرار گیرد. با پیشرفت سخت‌افزار کوانتومی و افزایش تعداد کیوبیت‌ها، انتظار می‌رود که مزایای محاسبات کوانتومی در این حوزه بیشتر آشکار شود.

مراجع

[۱] Lau, B., Emani, P.S., Chapman, J., et al. (2023). Insights from incorporating quantum computing into drug design workflows. *Bioinformatics*، 39(1)، btac789.

[۲] Farhi, E., & Neven, H. (2018). Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002*.

[۳] Qiskit Development Team. (2023). Qiskit: An open-source framework for quantum computing. https://qiskit.org/

[۴] Wang, Y., et al. (2019). GenoDock: A framework for predicting mutation impact on drug binding. *Bioinformatics*.