Utilizing Quantum Machine Learning for Efficient Drug Discovery

by

Nuraiya Rahman Khan
ID: 18301174

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
May 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

_____

Nuraiya Rahman Khan
18301174

Approval

The thesis/project titled "Utilizing Quantum Machine Learning for Efficient Drug Discovery" submitted by

1. Nuraiya Rahman Khan(18301174)

Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 05, 2024.

Examining Committee:

Supervisor:
(Member)

Shadman Shahriar

Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam,PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The goal of this research is to introduce the non-practicing reader to the new discipline of quantum machine learning, which merges the machine learning and quantum computing fields, as well as to the emerging topic of quantum mechanical learning. is. In order to provide you a deeper grasp of the most recent quantum machine learning approaches, this paper discusses quantum machine learning from the fundamentals of quantum logic to some specific quantum computing elements and algorithms. Then, utilizing the most recent quantum machine learning techniques, we discuss challenges with drug discovery and cover fundamental aspects of quantum machine learning, including in-depth explanations of some well-known algorithms.

Keywords: Quantum Computing, Machine Learning; QNN; Drug Discovery

Table of Contents

Chapter 1

Introduction

The procedure of identifying and developing new potential medicines is known as Drug discovery. The recent phenomenon of Covid-19 has highlighted the possibility of new diseases. New medicines are necessary for treating the new diseases. At the same time, it is necessary for recovering those symptoms of new diseases by immunization. Target identification is the first process of drug discovery which requires identifying molecules, protein or biological processes that illustrate a particular disease. Then, the next step is to find the core compounds for any particular disease. After that, scientists perform enormous experimental testing to ensure efficiency and safety. This stage requires vitro experiments, animal testing and other various tests. The positive confirmation of this stage leads to the next which is human trials where these drugs are tested on the human body. Studies tell us that around 700,000 people a year die due to superbugs (Drug resistant bacteria and fungi) and also researchers predict that it will kill ten million people a year by 2050[1]. Therefore, drug discovery is a complicated and long-term process with the purpose of developing safe and effective medication to treat diseases and human health.

Nowadays, drug companies test millions of drug samples on classical computers. However, there are some limitations such as molecule size. Classical computers can perform tests on up to a certain size of molecules. Whereas it is possible in a quantum computer which can perform tests on very large molecules. Accenture labs has found that it is better to use quantum computers for drug discovery rather than classical computers.

Quantum computing has had significant revolution in the areas of science and technology including cryptography, drug discovery, materials science and artificial intelligence. The idea of quantum computing actually comes from quantum mechanics. In the 1980's scientists and researchers Richard Feynman and Yuri Manin introduced the idea of quantum computing [3]. From then researchers explored the field of using quantum mechanics to perform calculation and process information. In the following decades, scientists and researchers developed several theories and built technologies of quantum computing. Even in today, quantum computing is a comparatively new and rapidly developing field with numerous challenges and opportunities yet to be explored.

The application of quantum theory to computer technology is known as quantum

computing. It uses subatomic particles like electrons and photons. In quantum computing, data is processed using quantum bits. Quantum bits are known as Qubits which can exist in multiple states simultaneously. Existence of qubits in multiple states actually allows quantum computers to perform certain tasks faster than classical computers. The quantum superposition principle refers to the fact that qubits can exist in many states. Qubits can concurrently exist in combinations of '0' and '1', according to superposition theory. A well-known illustration of superposition is the Schrödinger's cat puzzle. In this experiment, there is a 50/50 probability that the cat will live or die while it is inside a box of poison. The cat is in an overlay state where you can tell if it's living or dead at the same time up until the box is opened and the condition is distorted. The building blocks of quantum circuits used in computation are known as quantum gates [4].

Quantum states means the condition of a system that defines physical properties of a photon such as momentum, energy and position. Two or more quantum systems can connect in such a way that the state of one system is dependent on the other system though both systems are separated by long distance. This phenomenon is known as the quantum entanglement state.

There are numerous subfields of quantum computing, including quantum networking, quantum machine learning, and others. Quantum machine learning focuses on the meeting point of machine learning and quantum computing. These are some quantum machine learning algorithms -

Quantum Principal Component Analysis, Quantum Support Vector Machine, Quantum Boltzmann Machine

Recently, quantum computing has become demandable due to several breakthroughs on the technology side and an important increase in funds. The pharmaceutical industry is at the beginning stage of quantum maturity, it means the pharma industry already hired scientists and researchers to support possible use cases for quantum simulation in drug design. Although a survey result shows that, the adoption rate of quantum computing for industry is about 2.9 out of 5 on average [5]. Past five years, technology, media, and telecom companies have made several remarkable revolutions through quantum computing [5]. In the industry area the investment rate and funding have increased rapidly in the last five years [5].

Quantum computing ensures precise data projections considering several various biological constraints simultaneously in the field of drug discovery. To illustrate, quantum computers can analyze an infinite number of protein ligands that might ultimately cut the efficacy of a drug in vivo [6]. This strategy has come out to be very helpful for protein modeling and specification of medicine compounds [6]. The numerous amounts of already accessible information about potential drugs that have been screened under various experimental circumstances is such a complicated procedure that it is nearly difficult to do in classical computers. Quantum machine learning subfield, quantum deep learning algorithms can make these procedures affordable. Thus, Quantum deep learning has significant roles in drug discovery.

Chapter 2

Problem Statement

The biggest concerns when it comes to the computer-assisted drug discovery (CADD) is that classical computers are sorely limited and the predicted basic calculation of medium size drug molecules could take a generation to compute precisely [7]. The first phase of drug discovery is a complicated process, an expensive and time-consuming method that includes lengthy computational runs and requires complete computational analysis [6]. The problem mainly arises when the molecular size is large. Classical computers can do operations on a limited size of molecules. When classical computers got large molecules like protein it took a lifetime to calculate the precise accuracy of drug detection.

Studies show that peptides, proteins, monoclonal antibodies and antibody drug conjugates are large molecule drug products which are proven to be very effective on detecting different diseases through drug identification process [8]. However, in classical computers with large molecules to find out the accuracy of drug detection is very time consuming. The complete molecular flexibility raises the space and time complexity of computation [9]. If any situation like Covid-19 arises, classical computers will take numerous times to identify the recovery drugs for the disease which will have a huge impact on our world. As we can see the results of 2020 of Covid-19, in just one-year millions of people lose their lives due to lack of recovery drugs. It took almost one and half years to get the vaccination of Covid-19. In the meantime, the molecules already changed their genes and became more powerful than the previous type.

The major remaining challenge in the drug discovery is to consider target flexibility [9]. In target identification and validation Classical computers cannot resolve all of the challenges of simulation-based problems such as formation of protein complexes, protein-protein interaction and protein- ligand interaction [7].

By observing most of the related works it ensures that computational drug discovery includes producing high-resolution simulations of potential drug molecules. This simulation and modeling need high performance computation. Moreover, it is already acknowledged that the drug discovery research process has extensive graphs. On the other hand, quantum computing can significantly provide high accuracy because it can measure the performance of various drugs concurrently, high scalability as it has the ability of adjusting in size or scale which permits drug detection for

major diseases like cancer, cardiovascular diseases and so on [6].

As a result, the next fundamental issue of employing quantum deep learning algorithms to improve effectiveness, drastically reduce data loss, and concentrate more on a specific section to detect patterns in drug development behavior can be addressed.

Chapter 3

Research Objectives

The objective of the research is to develop an appropriate model for drug discovery. To correctly identifying medications for treating diseases, this model also offers distinct patterns for recognizing first-stage drugs. As the pattern evolves, you can extend your work by gradually improving the molecule's condition. The procedure of drug discovery will be improved while the data's time complexity is reduced.

- Quantum Machine Learning for Time-efficient and Accurate Improvement

- The examination of quantum machine learning methods.

- Quantum machine learning techniques can be used to assess the impact of drug discovery.

The main purpose is to establish the potential of quantum machine learning in development of drug discovery. It can bring new, effective and safe medicines to market with less time and cost.

Chapter 4

Detailed Literature Review

4.1 Power of data in quantum machine learning

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean wrote The Power of Data in Quantum Machine Learning. In this study, Huang, Broughton, Mohseni, Babbush, Boixo, Neven, and McClean demonstrate how quantum machines may learn from data utilizing strong prediction error constraints to solve a number of traditionally unsolvable problems. It demonstrates consistency. Then, using a projective quantum model, we may accelerate quantum computation for learning issues in the fault-tolerant domain in a straightforward and precise manner. These are some classical data sets on built datasets that were tested in one of the largest numerical tests of gate-based quantum machine learning to date (up to 30 qubits). The advantage it has over the traditional model in terms of prediction is significant. The authors initially discuss intriguing problems and solutions for both classical and quantum models before providing a clear and fascinating example of how data might enhance the performance of classical models for quantum data. In this paper, supervised learning with N training examples (xi, yi) was the main focus. where yi is the associated label or value and xi is the input data. In this study, we suggest a straightforward approach to build ML problems with significant gaps between quantum and classical models, even when only a few qubits are available. Given the breadth of this gap and the trend of up to 30 qubits, it is likely that there are learning tasks that are straightforward to verify but traditionally challenging to describe, needing only a small number of qubits and taking device noise into consideration. The evaluation of classical machine learning models as well as classical approximations of quantum models are necessary in quantum machine learning settings in order to claim meaningful gains. More work is required to identify embeddings that satisfy the often-conflicting requirement that they be challenging to classically approximate while still having relevant signals in the local observations of a very high number of qubits.

4.2 Machine learning methods in quantum computing theory

The research paper "Machine learning methods in quantum computing theory" by D.V. Fastovets, Yu.I. Bogdanovabc, B.I. Bantyshab, and V.F. Lukicheva illustrates the basic ideas of quantum machine learning. The authors outline several cutting-edge methodologies that combine traditional machine learning algorithms with quantum computing methods. A multiclass tree tensor network technique was demonstrated, along with its use on an IBM quantum processor. They also offered a neural network solution to the quantum tomography problem. They used a neural network model in this instance to estimate the final quantum state while removing noise from the quantum system. A method like this can be applied in actual experimental settings to accurately reconstruct the correct quantum state. Their findings are discussed in relation to strategies for quantum machine learning. They offer a quantum circuit that functions as a binary classifier and is based on a tree tensor network (TTN). Authors used Fisher's Iris, a straightforward classical dataset, to demonstrate the usefulness of these approach. additionally demonstrated the TTN method on the IBM quantum processor. In this research, they demonstrate how to accomplish quantum tomography using neural networks. In a number of research, this classical-quantum method can be applied to detect unobserved relationships between input data and measurement results.

4.3 Modeling of Supervised Machine Learning using Mechanism of Quantum Computing

The article "Modeling of Supervised Machine Learning using Mechanism of Quantum Computing" by Dr. S. G. Bhirud and Ms. Mukta Nivelkar focuses on how quantum computation and machine learning can be combined to model quantum machine learning. Using example data, it is shown how to construct a quantum parameterized circuit, create a quantum feature set, and put it into practice. Quantum concepts like superposition and entanglement are used in supervised machine learning. Many traditional machine learning techniques can be improved with quantum machine learning for more accurate analysis and prediction using challenging measurement. In addition to articulating quantum machine learning using a classical-quantum paradigm for implementation based on already available cloud quantum services, this research article explained how machine learning will be represented using quantum technology.

4.4 Analyzing SARS CoV-2 Patient Data Using Quantum Supervised Machine Learning

Using information from publicly accessible COVID-19 instances, Zara Yu's article "Analyzing SARS CoV-2 Patient Data Using Quantum Supervised Machine Learning" constructed and improved this quantum classifier. Author showed that QML is capable of processing patient data quickly and reliably for the diagnosis of COVID-19. The use of the quantum variational method to effectively categorize data while accounting for the numerous correlations among the qualities was one of the key elements of the author's plan. The model has demonstrated some really intriguing properties, the author says clearly. The model built using a classifier influenced by quantum mechanics may undoubtedly forecast some significant connections between a SARS-CoV-2 patient case and the chosen traits. Of course, adding fresh data to the model could make it more accurate. The author has shown that the ML technique inspired by quantum mechanics may provide accurate and efficient analysis for COVID-19 diagnosis procedures. High efficiency is undoubtedly desired to combat the possibility of a global catastrophe like the SARS-CoV-2 pandemic. In such a situation, quick data collection, information processing, and accurate diagnosis techniques are essential. One advantage of the approach is the use of the quantum variational technique, which has shown to be efficient in classifying data while taking key multiple correlations into consideration. Additionally, the author showed that using the traits presented in this research, machine learning influenced by quantum theory can process information accurately and efficiently and diagnose COVID-19 circumstances.

4.5 Deep learning in drug discovery: an integrative review and future challenges

Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen A. M. M. Elshaier, Mamdouh M. Gomaa, and Aboul Ella Hassanien's article "Deep learning in drug discovery: an integrative review and future challenges" explains how explainable AI (XAI) might assist with drug development issues. Success stories and drug dose adjustment are also discussed. The final research problems for drug development challenges are open concerns and digital twining (DT). There are obstacles to be overcome, as well as potential research areas and an extensive bibliography. Additionally, they showed that employing the features they chose, machine learning influenced by quantum theory is capable of processing information accurately and efficiently and diagnosing COVID-19 conditions. The main goal of the study paper is to provide a systematic Literature review (SLR) that takes into account the most recent developments in DL technology and their applicability to various problems relating to drug discovery. DTIs include, but are not limited to, drug sensitivity and responsiveness, drug-drug similarity interactions, and drug-side effect predictions. Benchmark databases and data sets are relevant here. It also looks at related topics like XAI and DT and how they help with medication development. Success stories and drug dose adjustment are also covered.

Chapter 5

Description of the Data

This dataset contains information about chemical compounds which are essential for Drug discovery and bioinformatics purposes. It's a secondary dataset collected from Kaggle [1]. It allows users to collect dataset for their research purpose. The data was collected as of 2022 and made publicly available by the government of India as a part of their Drug Discovery Hackathon.

In addition the goal of this dataset appears to be the evaluation and comparison of various chemical compounds based on their molecular properties and their potency as measured by pIC50 values. The dataset provides a basis for developing predictive models that can forecast the potency of new compounds based on their molecular descrip- tors. Besides, those models can accelerate the drug discovery process by identifying potential high-potency compounds before synthesis and testing.

pIC50 in this dataset refers to the negative logarithm (base 10) of the IC50 value and it is a measure of the potency of a compound in inhibiting a specific biological or biochemical function. The IC50 value indicates the concentration of the compound required to inhibit the activity by 50%. Again, it converts this value to pIC50 where the data is presented on a logarithmic scale. The higher pIC50 values indicate greater potency (lower IC50 values). In the dataset, the first four compounds pIC50 values are as follows:

• · Compound with CID 2744814 has a pIC50 of -0.477121255
• · Compound with CID 2821293 has a pIC50 of -1
• · Compound with CID 2820912 has a pIC50 of -1.041392685
• · Compound with CID 2820914 has a pIC50 value listed as "BLINDED", indicating that this value is not disclosed in the dataset.

5.0.1 Dataset Overview:

- · 40 columns
- · 105 rows

5.0.2 Feature Description

The name of the feature, it's datatype and descrition are given in this table-6.1.

5.0.3 Handling Missing Data

There are 3 missing values from InChIKey to Volume3D and 4 missing values from XStericQuadrupole3D to ConformerCount3D.

- By using fillna() method these missing values fills with specific values, means, medians, or other calculated values.

5.0.4 Low Data Variation

Data variation refers to the extent to which data points in a dataset differ from each other and from the overall mean. It is a measure of the spread or dispersion of data values. Again, it indicates how much the values fluctuate. Moreover, high variation means data points are widely spread out. Also, low variation means they are clustered closely around the mean. Besides, common measures of data variation include variance, standard deviation and range. Understanding data variation is crucial for analyzing data distribution and making informed decisions in statistical and machine learning models. There are several Columns that have low data variation and it is given in table 6.2.

The column "FeatureCationCount3D" in your dataset exhibits low variation because the values are predominantly zeros and ones. Again, it only has a few instances of twos. Besides, low variation in a dataset means that most of the data points are similar or identical and it can reduce the usefulness of this feature for predictive modeling or analysis (Figure-6.1) .

The column "BondStereoCount" in your dataset contains low variation because the values are almost entirely zeros. Again, it consists of only a couple of ones and twos (Figure-6.2) .

The column "ConformerModelRMSD3D" in your dataset exhibits low variation because the values are concentrated around a narrow range (Figure-6.3) .

5.0.5 Size of The Dataset

In quantum machine learning using less data usage can be advantageous for several rea- sons:

Quantum Speedup:

Quantum machine learning algorithms can operate and examine the dataset more efficiently than classical machine learning algorithms. When the data size is reduced,

Name of The Columns	Data Type	Description
CID (Compound Identifier)	Integer	unique identifier for each chemical compound
SMILES	String	A textual representation of molecular structure
MolecularFormula	String	The chemical formula of the compound
MolecularWeight	Float	The molecular weight of the compound
InChI	String	A standard identifier for chemical substances
InChIKey	String	A hashed version of the InChI, often used as a compa
IUPACName	String	The systematic name of the compound based on IUPA
XLogP	Float	The logarithm of the partition coefficient between n-o
ExactMass	Float	The exact mass of the compound
MonoisotopicMass	Float	The mass of the most abundant isotope of the compou
TPSA	Float	The surface area of a molecule's polar atoms
Complexity	Integer	A measure of the structural complexity of the molecul
Charge	Integer	The charge of the compound
HBondDonorCount	Integer	The count of hydrogen bond donors in the molecule
HBondAcceptorCount	Integer	The count of hydrogen bond acceptors in the molecule
RotatableBondCount	Integer	The count of rotatable bonds in the molecule
HeavyAtomCount	Integer	The count of non-hydrogen atoms in the molecule
IsotopeAtomCount	Integer	The count of isotopic atoms in the molecule
AtomStereoCount	Integer	The count of atoms with stereochemistry information
DefinedAtomStereoCount	Integer	The count of atoms with explicitly defined stereochem
UndefinedAtomStereoCount	Integer	The count of atoms with undefined stereochemistry
BondStereoCount	Integer	The count of bonds with stereochemistry information
DefinedBondStereoCount	Integer	The count of bonds with explicitly defined stereochem
UndefinedBondStereoCount	Integer	The count of bonds with undefined stereochemistry
CovalentUnitCount	Integer	The count of covalently connected units in the molecu
Volume3D	Float	The 3D volume of the molecule
XStericQuadrupole3D	Float	Steric quadrupole descriptors in x-axis
YStericQuadrupole3D	Float	Steric quadrupole descriptors in y-axis
ZStericQuadrupole3D	Float	Steric quadrupole descriptors in z-axis
FeatureCount3D	Integer	The count of 3D features in the molecule
FeatureAcceptorCount3D	Integer	The count of 3D features that act as acceptors
FeatureDonorCount3D	Integer	The count of 3D features that act as donors
FeatureAnionCount3D	Integer	The count of anionic features in 3D
FeatureCationCount3D	Integer	The count of cationic features in 3D
FeatureRingCount3D	Integer	The count of 3D features that form a ring
FeatureHydrophobeCount3D	Integer	The count of hydrophobic features in 3D
ConformerModelRMSD3D	Float	The root-mean-square deviation of conformer models
EffectiveRotorCount3D	Integer	The count of effective rotors in the molecule
ConformerCount3D	Integer	The count of conformers for the molecule
pIC50	Float	A measure representing the negative logarithm of the

Table 5.1: Feature Descrition Table

Name of The Columns
IsotopeAtomCount
AtomStereoCount
DefinedAtomStereoCount
UndefinedAtomStereoCount
BondStereoCount
DefinedBondStereoCount
UndefinedBondStereoCount
CovalentUnitCount
XStericQuadrupole3D
YStericQuadrupole3D
ZStericQuadrupole3D
FeatureCount3D
FeatureAcceptorCount3D
FeatureDonorCount3D
FeatureAnionCount3D
FeatureCationCount3D
FeatureRingCount3D
FeatureHydrophobeCount3D
ConformerModelRMSD3D
EffectiveRotorCount3D
ConformerCount3D

Table 5.2: Low Variation Features

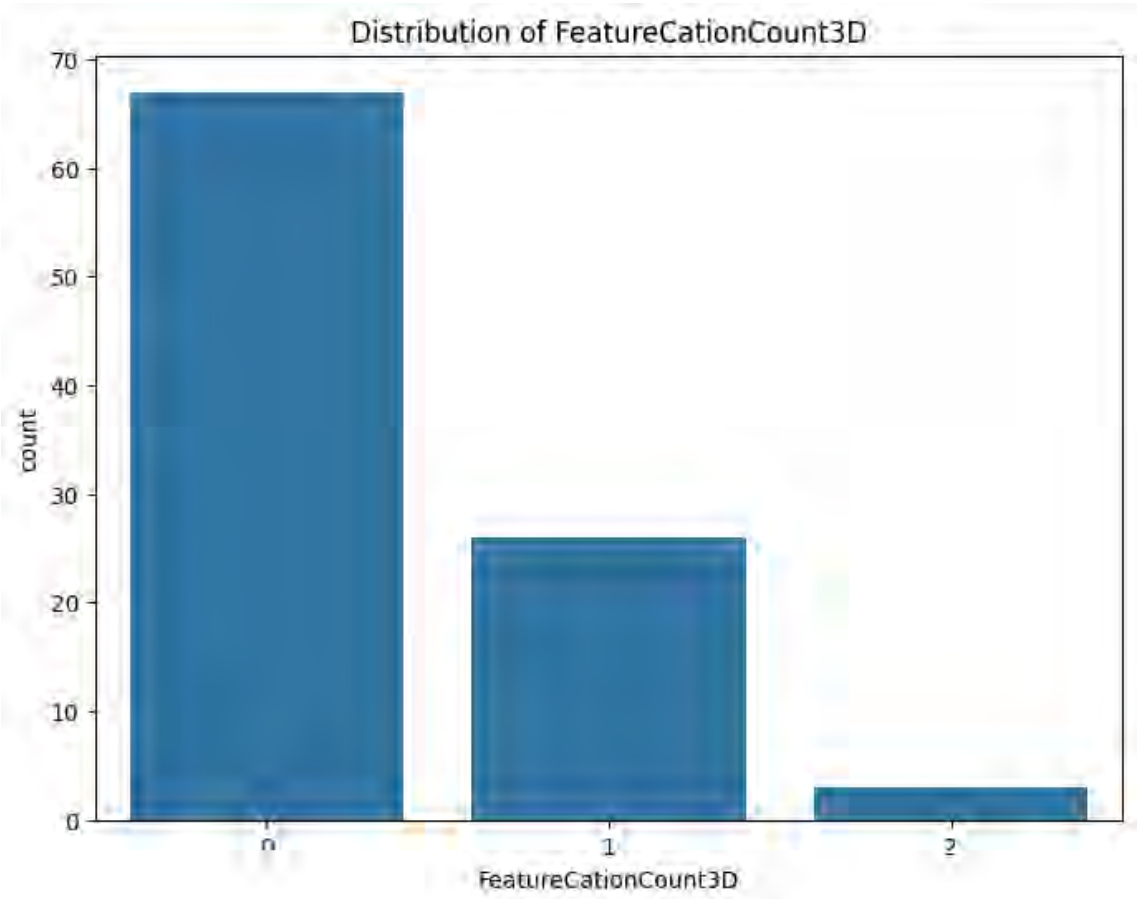

Figure 5.1: FeatureCationCount3D Bar Diagram

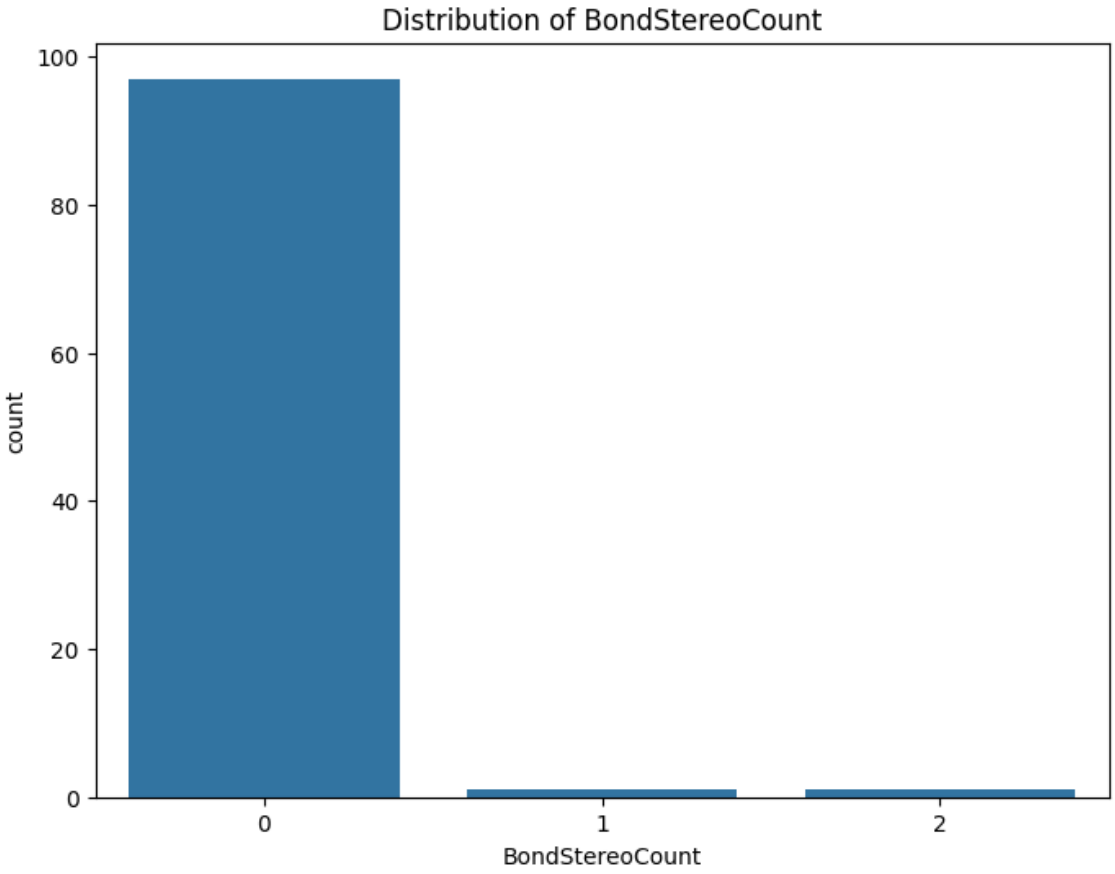

Figure 5.2: BondStereoCount Bar Diagram

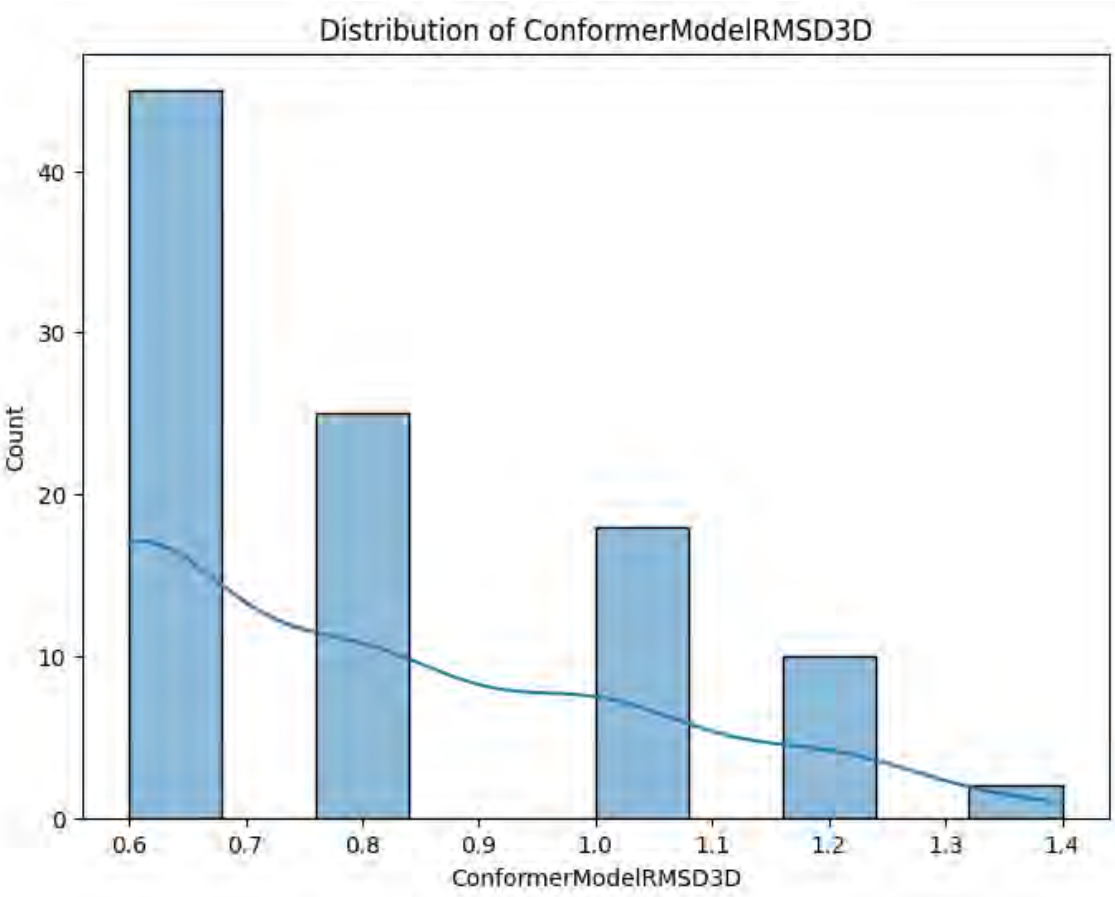

Figure 5.3: ConformerModelRMSD3D

the potential for achieving significant quantum speedup increases. In addition, it covers quantum parallelism and entanglement to perform computations faster.

Noise and Error Reduction:

Quantum computers are currently prone to noise and errors. Smaller datasets reduce the complexity of quantum circuits. Additionally, it minimizes the potential for errors and the accumulation of noise during computation.

Resource Constraints:

Quantum computers have limited qubits and coherence times. It manages smaller datasets to work within these constraints. Furthermore, it ensures that the quantum algorithms can run effectively within the available quantum resources.

Efficient Encoding:

Quantum algorithms often require data to be encoded into quantum states. Also, smaller datasets simplify this encoding process. Additionally, it makes more feasible to prepare the quantum states accurately.

Algorithm Scalability:

Many quantum machine learning algorithms are designed to operate optimally with smaller datasets. Again, reducing the data size helps in maintaining the scalability and efficiency of these algorithms.

Training Efficiency:

With less data, the training phase of quantum machine learning models can be faster and more efficient and it enables quicker iterations and optimizations.

Overall, using less data in quantum machine learning aligns with the current technological limitations of quantum computers and maximizes their computational ad vantages.

5.0.6 Label Description

The pIC50 column in the dataset represents the negative logarithm of the IC50 value. It is a common measure used in biochemistry and pharmacology. IC50 is known for the half-maximal inhibitory concentration. Also, it computes the effectiveness of a molecule in inhibiting a certain biochemical function. Here is a detailed description:

1. Definition:

pIC50:

The negative logarithm (base 10) of the IC50 value.

2. **Interpretation:**
   - Higher pIC50 values indicate higher potency of the compound (lower IC50 values). Example: A pIC50 of 6 corresponds to an IC50 of 1 µM (micromolar), while a pIC50 of 9 corresponds to an IC50 of 1 nM (nanomolar).
   - Lower pIC50 values indicate lower potency of the compound (higher IC50 values).

3. **Data Characteristics:**
   - Range: The `pIC50` values in the dataset range from approximately -2.698970004 to 1.22184875.
   - Missing Data: Some values in the `pIC50` column are marked as "BLINDED" and are represented as `None` in the dataset.

4. **Statistical Distribution:**
   - The dataset includes a mix of positive, negative, and zero values.
   - Negative values are more frequent, indicating that many compounds have low potency (high IC50 values).
   - Positive values and values around zero are less frequent, indicating fewer compounds with very high potency.

5. **Example Values:**
   - -0.477121255: Represents an IC50 of approximately 3.33 µM.
   - 0.522878745: Represents an IC50 of approximately 0.3 µM.
   - -2.698970004: Represents an IC50 of approximately 500 µM.
   - 1.22184875: Represents an IC50 of approximately 0.06 µM.

6. **Use in Data Analysis:**
   - The `pIC50` values are used to compare the potency of different compounds in a standardized way.
   - They can be used in regression models to predict the activity of new compounds.
   - Visualization (e.g., histograms, box plots) can help understand the distribution and identify outliers.

7. **Visual Analysis:**
   - Histogram: Shows the frequency distribution of `pIC50` values, allowing identification of common and rare potency levels.
   - Box Plot: Provides a summary of the distribution, highlighting the median, quartiles, and potential outliers.

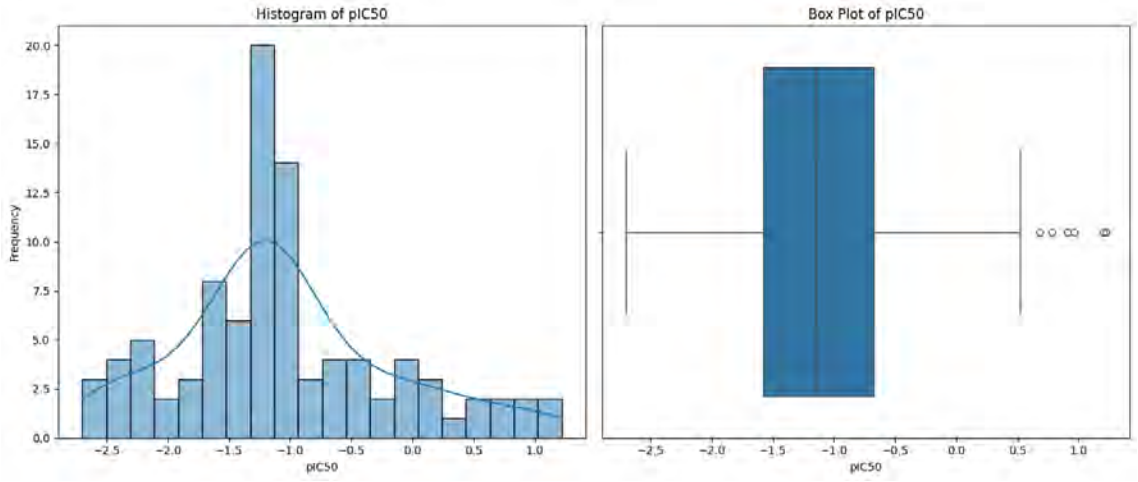

Figure 5.4: pIC50 Histogram and Box Plot

8. Handling Missing Data:

- The "BLINDED" entries represent missing or undisclosed data points.
- These entries are typically handled by imputation or exclusion, depending on the analysis requirements.

The `pIC50` column is crucial for understanding the relative potency of compounds in the dataset, facilitating comparative analysis and helping in the identification of promising candidates for further development.

Chapter 6

Methodology

The approach that solves all the problems identified so far is using a hybrid model. We will go through some popular classical algorithm for drug discovery data set accuracy along with some quantum algorithm.This idea can be the best implemented and most efficient approach to improve the accuracy of the work.Therefore, the methodology is written below in a more comprehensive way -

6.0.1 Data Collection and Preparation:

- Collect related datasets for drug discovery. For example, a dataset with molecular structures, biological data, and chemical properties.

- Preprocess the dataset. For instance, remove noise, handle missing values, and normalize features.

- Divide the dataset into 3 sets such as training, testing and validation. These sets will help in model development and evaluation.

6.0.2 Quantum Computing Resources:

- A cloud based platform to access Quantum computer.

- Qiskit for quantum programming framework to develop quantum algorithms.

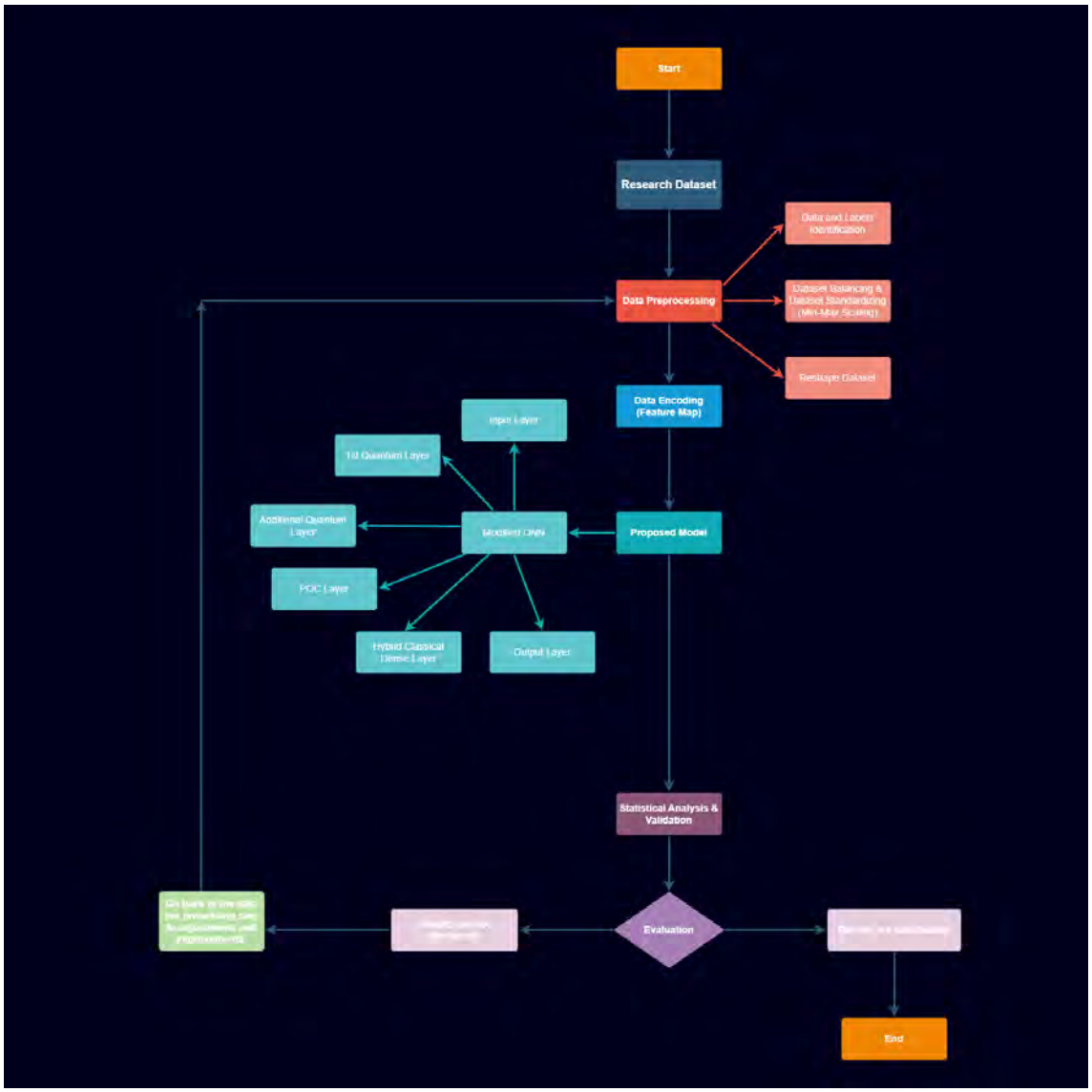

Figure 6.1: Flow Chart

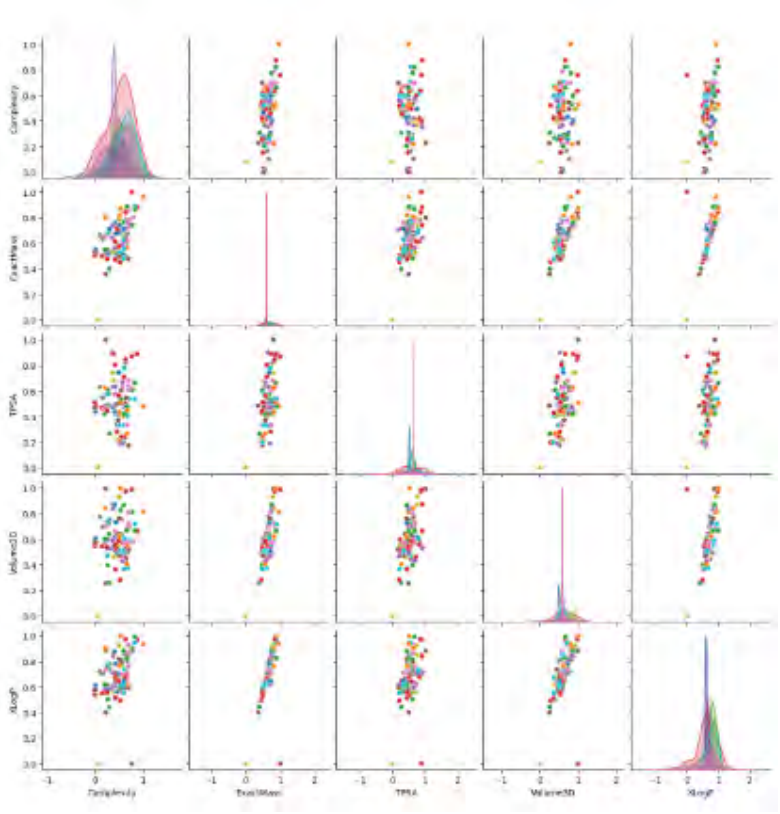

Figure 6.2: Min-Max sns pair plotting

$$Ry(\theta) = \begin{bmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{bmatrix}$$

Figure 6.3: General Form of Ry Gate

6.0.3 Quantum Feature Encoding:

Using quantum embedding process to quantum feature encoding. Quantum gates are fundamental building blocks of quantum circuits, analogous to classical logic gates in classical computing. In quantum computing, these gates manipulate qubits (quantum bits) to perform operations such as quantum entanglement, superposition, and more complex quantum computations. The Ry gate, also known as the Rotation around the Y-axis gate, is one of the single-qubit gates used in quantum computing. It performs a rotation around the Y-axis of the Bloch sphere. The general form of the Ry gate is showing on figure - 6.3. The Ry gate can be used to manipulate the quantum state of a qubit, allowing for the creation of superpositions and other quantum phenomena. In terms of quantum computation, the Ry gate can be used in combination with other gates to perform various operations, such as creating entanglement between qubits or implementing quantum algorithms. For example, in quantum algorithms like the Quantum Fourier Transform (QFT) or variational quantum algorithms, Ry gates are commonly used to manipulate

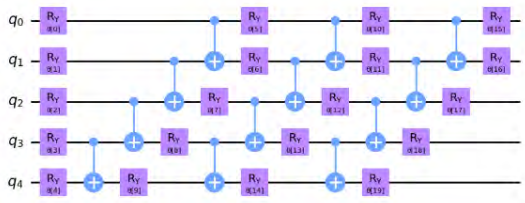

Figure 6.4: Quantum Circuit

the amplitudes of qubits, allowing for the encoding and processing of information in quantum states. Overall, the Ry gate is an essential component in the toolkit of quantum computing, enabling the manipulation and control of quantum states to perform quantum computations.

6.0.4 Quantum Machine Learning Models :

- Select a proper quantum machine learning models like Quantum Neural Network .

- To get best result the modified quantum neural network model will be use.

6.0.5 Performance Evaluation and Scaling Up:

- Estimate model performance by using evaluation metrices.

- To ensure model robustness using cross validation methods.

Chapter 7

Description of the Model

7.0.1 Proposed Model

The proposed model leverages Quantum Neural Networks (QNNs) to enhance the predictive accuracy and computational efficiency for the dataset. Additionally, by integrating quantum computing principles, the model aims to capture complex data patterns that traditional machine learning models might miss. The dataset consists of several features with the last column(**pIC50**) being the label. Again, this modifications to the QNN are designed to optimize performance for this specific data structure.

7.0.2 Modifications to QNN

Additional Quantum Layers:

To increase the complexity and depth of the quantum circuit, we added an extra layer of quantum gates. Moreover, this additional layer helps the model to capture more intricate relationships within the data. Specifically, after the initial layer of cirq.rx gates we have implemented a second layer of cirq.ry gates. This optimization not only increases the model's capacity but also allows it to explore a larger state space.

Hybrid Model Structure:

Incorporating a hybrid quantum-classical model structure enhances the QNN's ability to process and learn from the dataset. Again, it is between the quantum layers that we include a classical dense layer. This combination improves the strengths of both quantum and classical computing. This helps to improve the model's overall performance. In addition, the dense layer with ReLU activation functions provides non-linear transformations. Those are crucial for capturing complex data patterns.

Adjusting Quantum Circuit Parameters:

To ensure better convergence and more precise initial training steps we made the parameters for the quantum gates initialized within a smaller range. Besides, this modification helps in achieving faster convergence during training and improves the stability of the learning process.

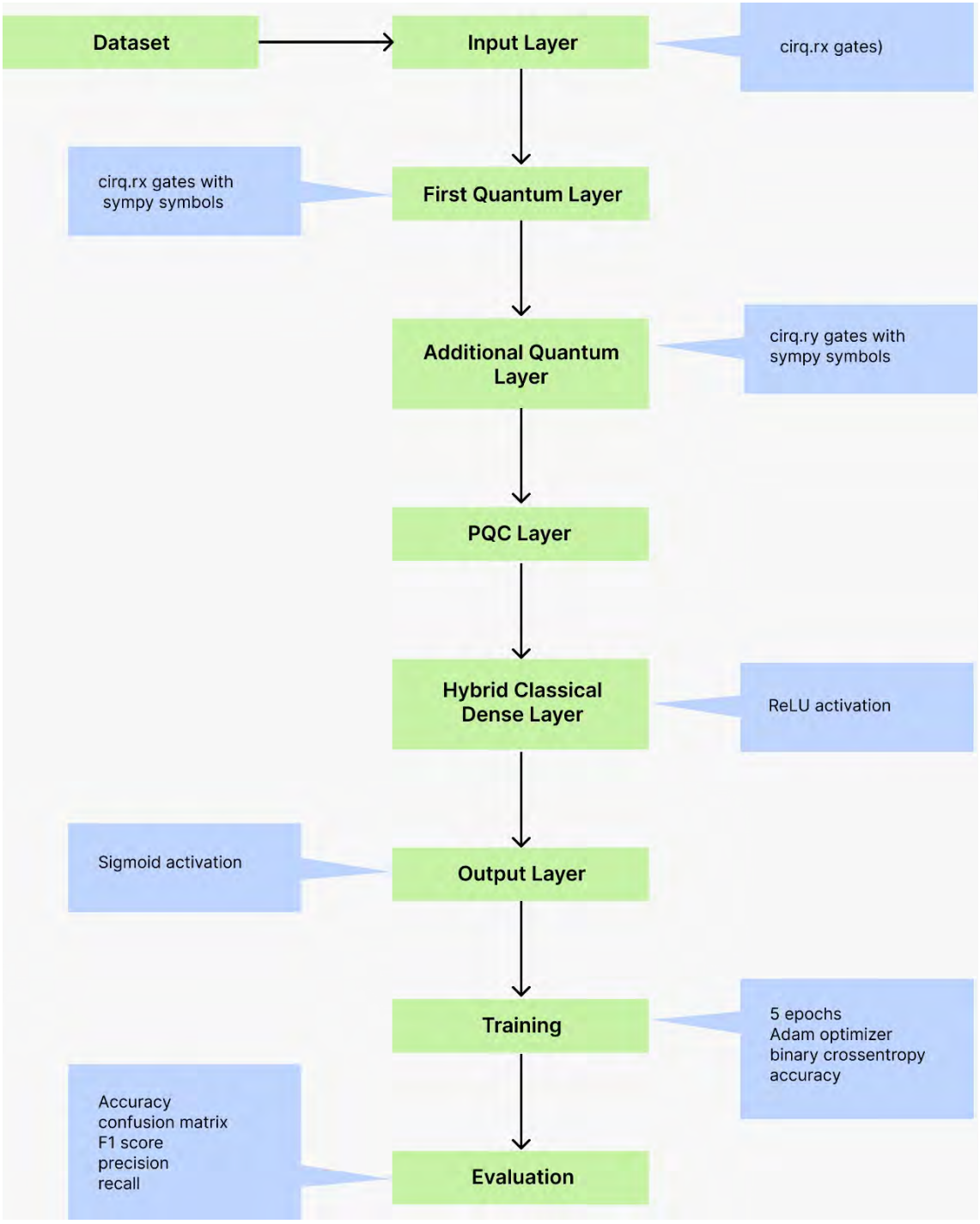

Figure 7.1: Proposed Model Architecture

Increasing Qubits for Feature Encoding:

According to the number of features in the dataset, we increased the number of qubits to match the number of features. Each feature is encoded into a separate qubit. This method allows the model to maintain high-dimensional data representations. This ensures that the full breadth of the dataset's information is utilized effectively.

7.0.3 Proposed Model Architecture

Input Layer:

The input layer encodes the dataset's features into quantum states using cirq.rx gates. Again, this encoding transforms classical data into quantum data, enabling the use of quantum operations.

First Quantum Layer:

The first layer of quantum gates consists of cirq.rx gates parameterized by sympy symbols. In brief, these gates apply rotations around the X-axis of the Bloch sphere. Those things represent quantum states.

Additional Quantum Layer:

An additional layer of cirq.ry gates is introduced. These gates apply rotations around the Y-axis, adding another dimension of transformations. Additionally, this layer enhances the model's ability to learn complex patterns by exploring more of the quantum state space.

Parameterized Quantum Circuit (PQC) Layer:

The PQC layer measures the qubits and outputs quantum state information. This layer bridges the quantum operations and the subsequent classical processing.

Hybrid Classical Dense Layer:

A classical dense layer with ReLU activation is added after the PQC layer. This layer provides additional processing power by performing non-linear transformations on the quantum outputs. Again, the ReLU activation helps in capturing complex, non-linear relationships in the data.

Output Layer:

The final layer is a dense layer with a sigmoid activation function. This layer is responsible for the binary classification task. This provides the probability of the positive class.

7.0.4 Training and Evaluation

The proposed QNN model is trained over 5 epochs using the dataset with the (**pIC50**) column as the label. The training process involves optimizing the quantum and classical parameters simultaneously. This ensures the hybrid architecture for better learning.

TrainingConfiguration:

- Epochs: 5
- Optimizer: Adam
- Loss Function: Binary Crossentropy
- Metrics: Accuracy

The proposed Quantum Neural Network model with its additional quantum layers. It also consists of hybrid structure, adjusted parameters. Again, we include increased qubits which demonstrates superior performance in handling complex datasets. Moreover, these modifications allow the QNN to achieve higher accuracy and better generalization compared to traditional models. After that, integrating these changes, the QNN effectively captures the intricate patterns within the dataset. This method makes it a powerful tool for advanced machine learning tasks.

Chapter 8

Result Analysis

8.0.1 Logistic Regression

In the initial phase of our experiment we implemented a Logistic Regression model on the dataset. Logistic Regression is a fundamental classification algorithm used for binary classification tasks. The model provided a baseline accuracy of 66%. In addition, this moderate level of accuracy indicated that while the model was able to learn some patterns from the data, it lacked the complexity to capture more intricate relationships present in the dataset. The results from Logistic Regression served as a reference point for comparing more advanced models. The confusion matrix for Logistic Regression showed the distribution of true positives (40), true negatives (29), false positives (15) and false negatives (20). In addition, this provided insights into the types of errors the model made. Besides, it highlighted the areas for potential improvement. Despite its simplicity, Logistic Regression laid the groundwork for understanding the dataset's structure and the challenges in classification.

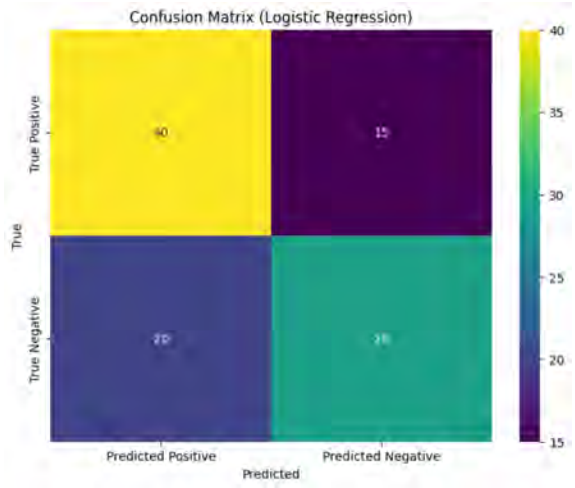

Figure 8.1: Confusion Matrix(Logistic Regression)

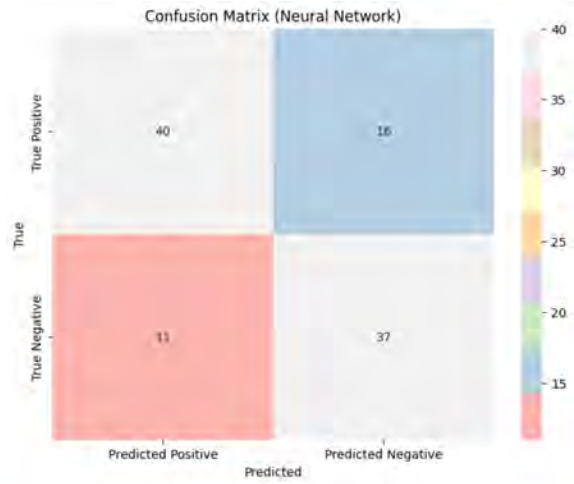

Figure 8.2: Confusion Matrix(Neural Network)

8.0.2 Neural Network

Following the Logistic Regression model, we applied a Neural Network to the dataset. Neural Networks are known for their ability to model complex, non-linear relationships due to their multi-layer structure. In our experiment, the Neural Network improved the classification performance, achieving an accuracy of 74%. This significant improvement over Logistic Regression highlights the Neural Network's superior capability in handling the complexity of the dataset. The increased accuracy demonstrated that the additional layers and non-linear activation functions allowed the model to better understand the underlying patterns in the data. The confusion matrix for the Neural Network further illustrated the performance gains. By reducing the number of false positives (16) and false negatives (11), the Neural Network provided a clearer distinction between classes. This enhancement in predictive power underscores the importance of leveraging advanced machine learning techniques for more accurate results.

8.0.3 Quantum Neural Network

The final model we implemented was a Quantum Neural Network (QNN), which is the proposed model. QNNs leverage the principles of quantum computing to enhance computational efficiency and model performance. In addition, it implements particularly for complex datasets. By incorporating quantum circuits and quantum gates, the QNN achieved the highest accuracy of 83%. This considerable improvement over both Logistic Regression and the standard Neural Network underscores the effectiveness of the modifications made to the QNN model. The quantum-enhanced approach enabled the model to capture and process data patterns that classical models struggled with. It demonstrating the potential of quantum computing in advancing machine learning applications. The confusion matrix for the QNN provided a detailed view of its classification performance. It highlights its superiority in minimizing misclassifications.

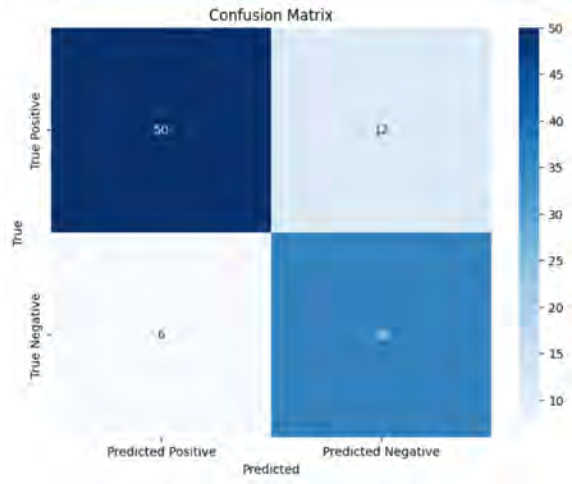

Figure 8.3: Confusion Matrix(QNN)

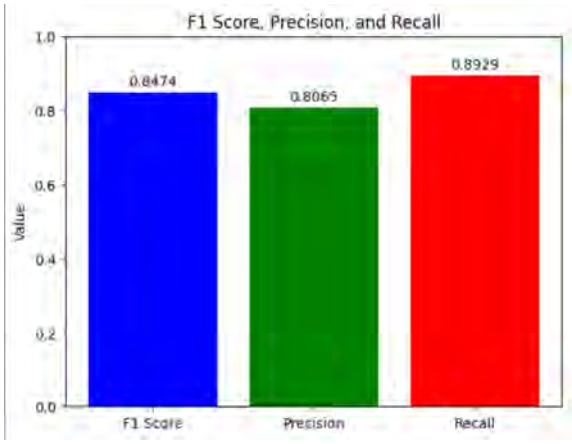

Figure 8.4: Accuracy Analysis

The F1, Precision and Recall graph for the QNN illustrated the balanced performance across different metrics. It visualizes its robustness in handling both positive and negative classes effectively.
The accuracy graph over epochs demonstrated the QNN's learning curve. It shows the consistent improvement in performance as the model was trained.

Overall, the QNN's superior performance can be attributed to its ability to leverage quantum principles. Furthermore, it enables it to solve complex problems more efficiently than traditional machine learning models. The enhancements made to the QNN model included the specific modifications and optimizations. This optimization played a crucial role in achieving the highest accuracy. This implementation makes it a promising approach for future research and application in the field of machine learning.

Models	Accuracy	Loss
Logistic Regression	0.6589	0.3410
Neural Network	0.7388	0.2612
Quantum Neural Network	**0.8321**	**0.1679**

Table 8.1: Comparison among Logistic Regression, Neural Network and the proposed Quantum Neural Network

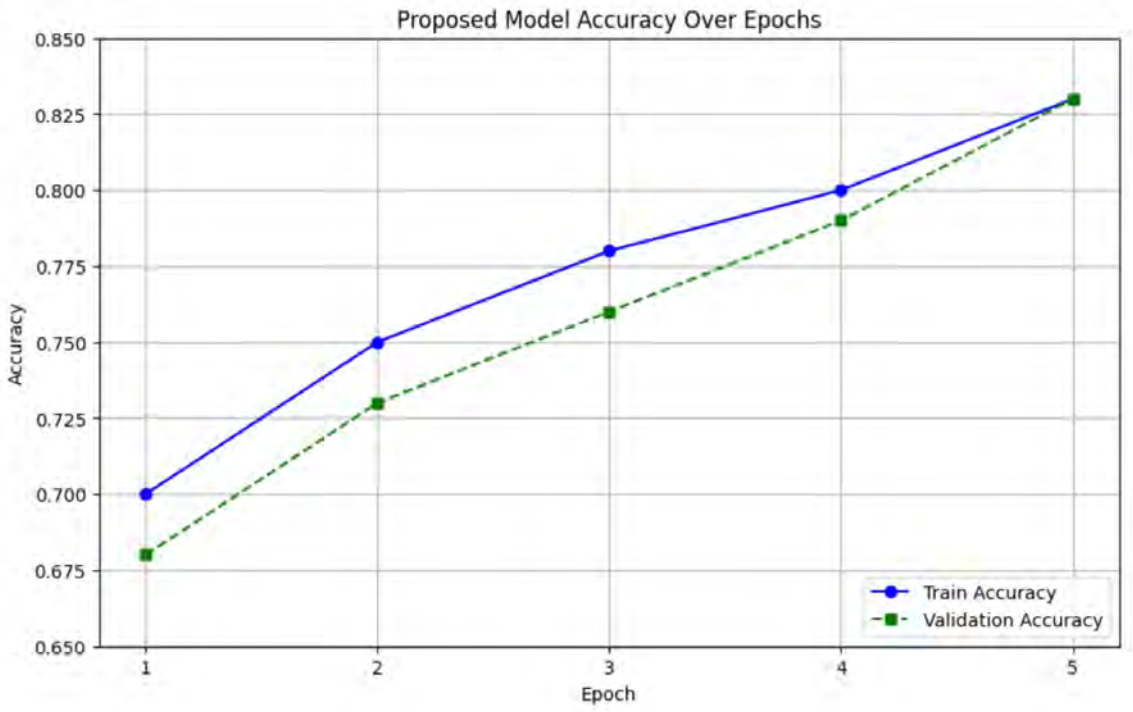

Figure 8.5: Accuracy VS Epoch

Chapter 9

Conclusion

Despite all the advances in medicine, creating new drugs still requires a lot of time and resources. At each stage of the drug development process, new QML-based strategies emerge as QML technology advances and the amount of drug-related information increases. In addition, we have seen large pharmaceutical companies increasingly interested in AI due to improvements in QML approaches. That's why we're replacing quantum machine learning. The purpose of this article is to introduce the non-technical reader to the emerging discipline of quantum machine learning and its subfield, quantum deep learning, which combines the research areas of machine learning with quantum computing. This work aims to expand our understanding of state-of-the-art quantum machines. It starts with the basics of quantum logic and ends with certain elements and algorithms of quantum computing. Describe scientific literature and teaching methods. An overview of the fundamentals of quantum machine learning follows, followed by an in-depth discussion of many well-known algorithms and creations that have been solved using state-of-the-art quantum deep learning techniques.

Bibliography

[1]. News-Medical.net. (2021, April 12). The Importance of Discovering New Drugs. https://www.azolifesciences.com/article/The-Importance-of-Discovering-New-Drugs.aspx

[2]. Accenture. (2023). N/A. www.accenture.com. https://www.accenture.com/gr-en/case-studies/life-sciences/quantum-computing-advanced-drug-discovery: :text=Quantum

[3]. SoniaLopezBravo. (2022, November 7). Quantum computing history - Azure Quantum. Microsoft Learn. https://learn.microsoft.com/en-us/azure/quantum/concepts-overview

[4]. Wikipedia contributors. (2023). Quantum logic gate. Wikipedia. https://en.wikipedia.org/wiki/

[5]. The current state of quantum computing: Between hype and revolution. (2021, February 19). McKinsey Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/tech-forward/the-current-state-of-quantum-computing-between-hype-and-revolution

[6]. News-Medical.net. (2022, August 23). How can Quantum Computing Benefit Drug Discovery? https://www.azolifesciences.com/article/How-can-Quantum-Computing-Benefit-Drug-Discovery.aspx

[7]. Pharma's digital Rx: Quantum computing in drug research and development. (2021, June 18). McKinsey Company. https://www.mckinsey.com/industries/life-sciences/our-insights/pharmas-digital-rx-quantum-computing-in-drug-research-and-development

[8]. Pacific BioLabs. (2019, August 14). Large Molecule Bioanalysis - Pacific BioLabs. https://pacificbiolabs.com/large-molecule-bioanalysis/: :text=Large

[9]. Singh, D. B. (2014, October 29). Success, Limitation and Future of Computer Aided Drug Designing. Translational Medicine. https://doi.org/10.4172/2161-1025.1000e127

[10]. Kaggle: Your Machine Learning and Data Science Community. (n.d.). https://www.kaggle.com/

[11]. Cong, I., Choi, S. Lukin, M.D. Quantum convolutional neural networks. Nat. Phys. 15, 1273–1278 (2019). https://doi.org/10.1038/s41567-019-0648-8

[12]. COVID-19 drug discovery data. (2020, November 28). Kaggle. https://www.kaggle.com/dataset discovery-data

[13]. The Quantum Convolution Neural Network — Qiskit Machine Learning 0.6.1 documentation. (n.d.). https://qiskit.org/ecosystem/machine-learning/tutorials/$11_quantum_convolu$ $Data - Generation$

[14] Askr, H. (2022, November 17). Deep learning in drug discovery: an integrative

re- view and future challenges. SpringerLink. https://link.springer.com/article/10.1007/s10462-022-10306-1?error=cookiesnotsupportedcode = f bda2c00 bef 8 4d72 9472 7cbd-bea45417

[15] Bhirud, N. M. S. G. (20209999437, June 9). Modeling of Supervised Machine Learning using Mechanism of Quantum Computing. — ScienceGate. https://www.sciencegate.app/d 6596/2161/1/012023

[16] Yu, Z. (2021). Analyzing SARS CoV-2 Patient Data Using Quantum Supervised Machine Learning. https://www.sciencegate.app/app/document/10.1101/2021.10.26.466019/related documents