GENOMIC COMPRESSION COMPARISSION WITH CNN,RNN AND WITH A UNIQUE APPROACH USING QCNN

22BIO211

TEAM 08

CH.SC.U4AIE23053– **Sri Hari Vasan A**
CH.SC.U4AIE23032 – **Mukesh Charan M**
CH.SC.U4AIE23061 – **Vigneshwarran S**
CH.SC.U4AIE23050 - Sanjjey S
CH.SC.U4AIE23047 - Rohit Mugalya A R

AMRITA VISHWA VIDYAPEETHAM | School of Engineering
AMRITAPURI | BENGALURU | CHENNAI | COIMBATORE

PROBLEM STATEMENT

TO USE QCNN(Quantum Convolution Nueral Network) for genome compression and we have compared it with the other models like normal CNN and RNN to show that QCNN has a better performance than that of the other two models

PREREQUSITES

Normal computation

head

or

Tail

Quantum computation

METHODOLOGY

- Used the quantum concepts like Quantum entanglement and superposition to reduce the memory space

Quantum Rotation Gates (RY Gates): Encode classical data into quantum states.

Controlled-NOT (CNOT) Gates: Introduce entanglement for better compression

Pauli-Z Expectation Measurement: Extract compressed data from quantum states.

HOW QCNN WORKS?

Qubit	Applied RY	Rotation (θ)
Qubit 1 (A)	RY(0.0)	No rotation
Qubit 2 (T)	RY(1.57)	Rotates by π/2
Qubit 3 (G)	RY(3.14)	Rotates by π
Qubit 4 (C)	RY(4.71)	Rotates by 3π/2

Encoding RY rotation = $|\psi\rangle = RY(\theta)\,|0\rangle = \cos(\theta/2)\,|0\rangle + \sin(\theta/2)\,|1\rangle$

Decoding

Pauli’s Z expectation

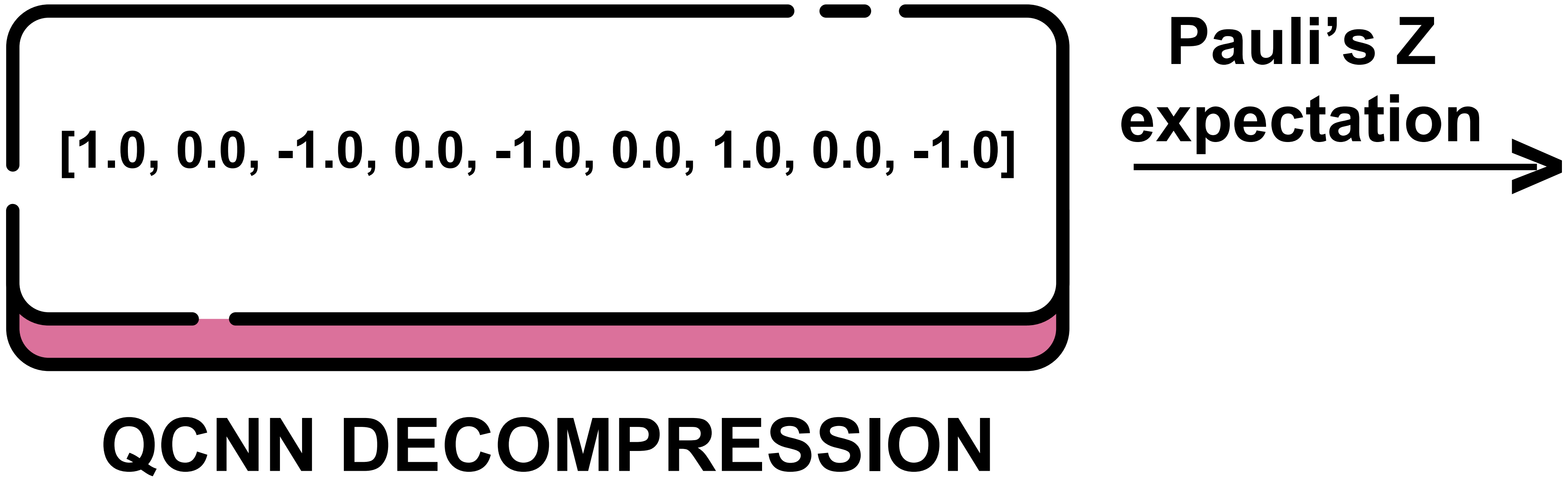
$\langle Z \rangle = \cos(\theta)$

ATCGCATTGAT.....

RY rotation

QCNN COMPRESSION

since all nucleotide can be stored at single QUANTUM STATE, the memory is reduced and computation time is faster

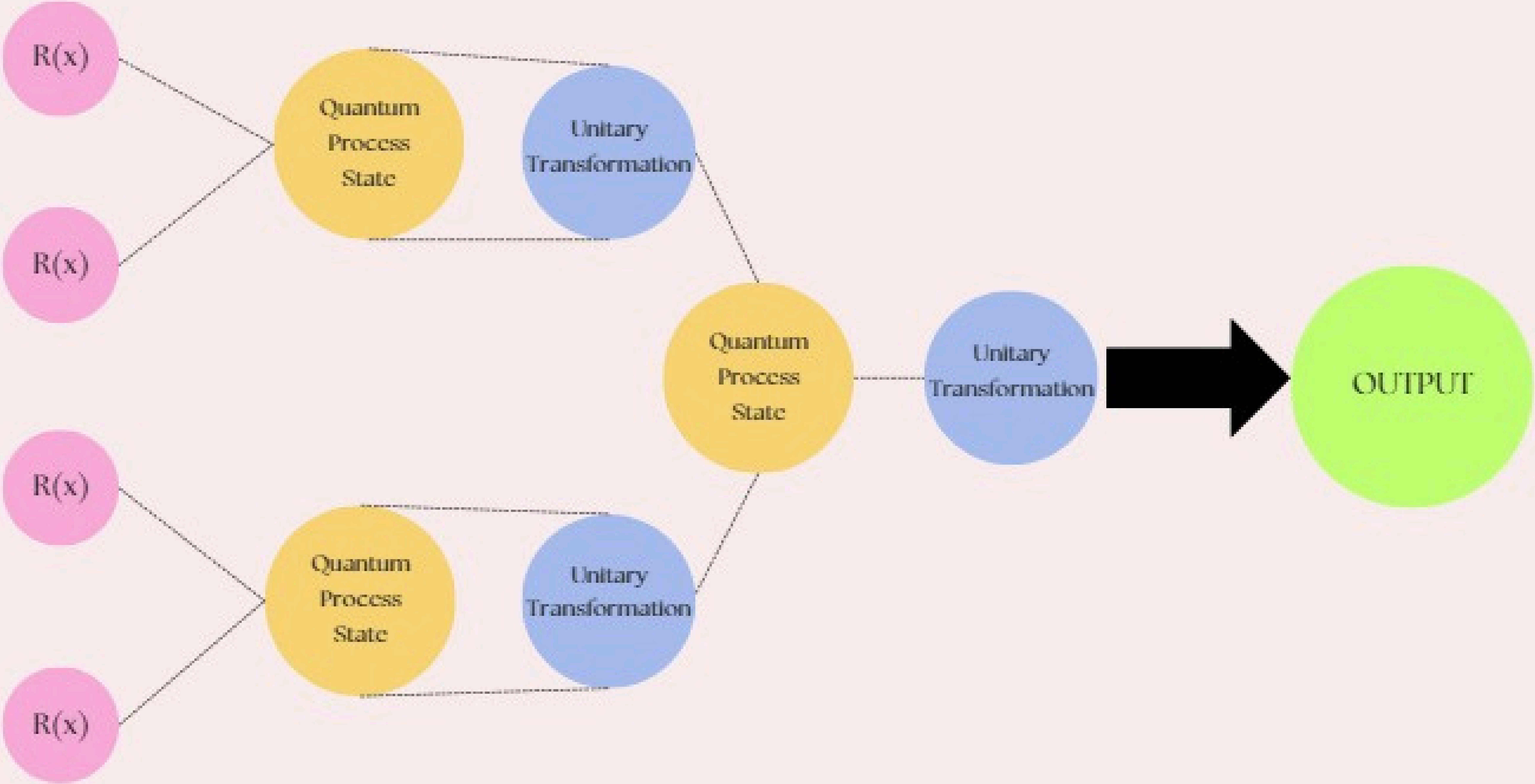


[1.0, 0.0, -1.0, 0.0, -1.0, 0.0, 1.0, 0.0, -1.0]

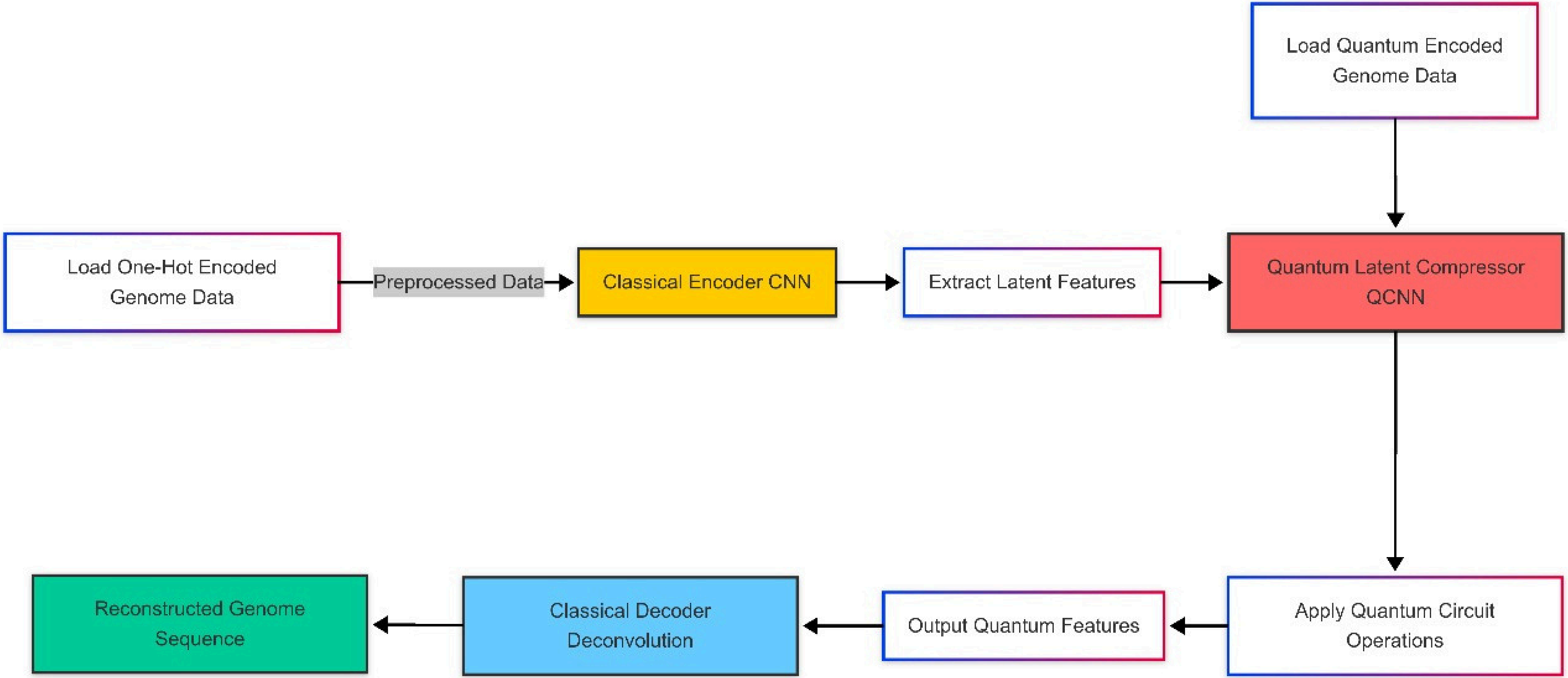
Pauli’s Z expectation

QCNN DECOMPRESSION

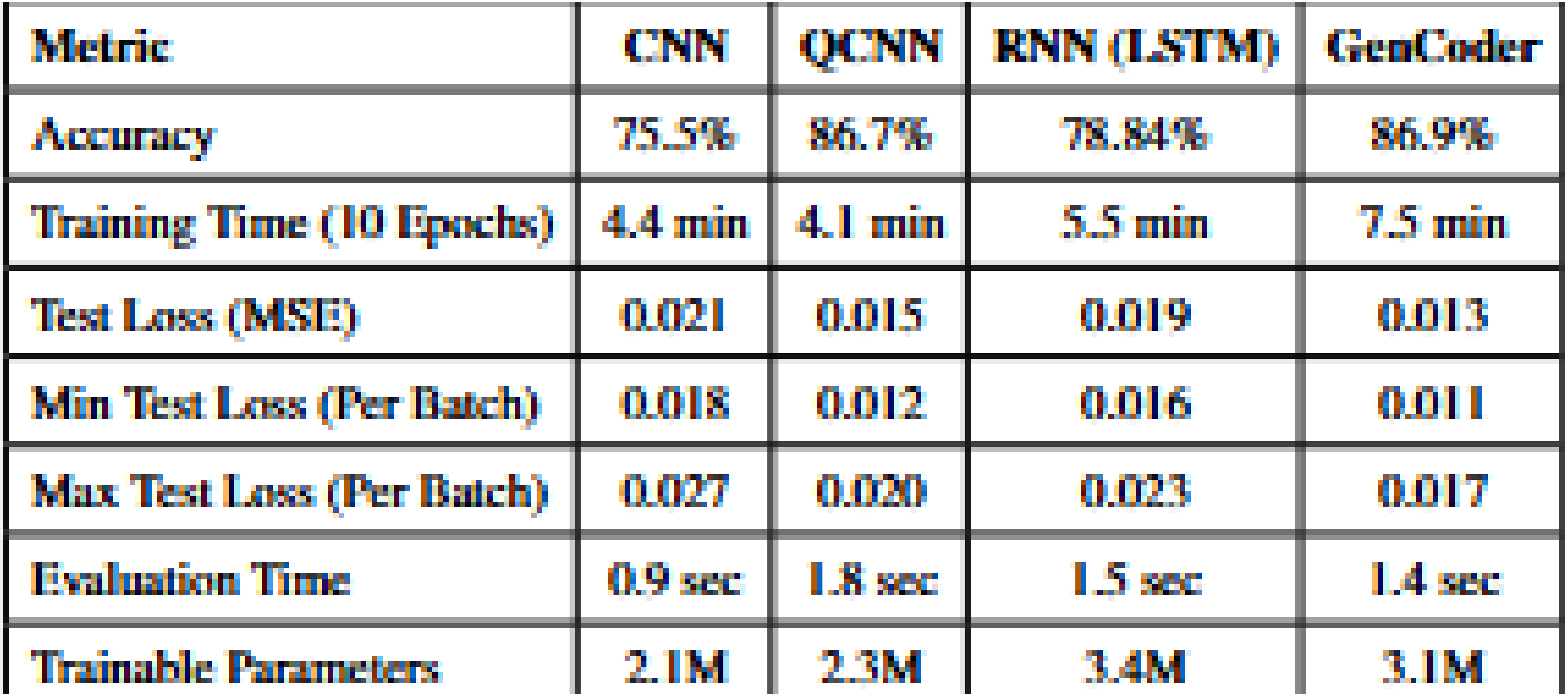
The Qubits are now restored back by using Pauli’s Z expectation for decoding purpose to prove the encoding is perfect


R(x)
R(x)
Quantum Process State
Unitary Transformation
Quantum Process State
Unitary Transformation
OUTPUT
R(x)
R(x)
Quantum Process State
Unitary Transformation

Technical approach


Load Quantum Encoded Genome Data
Load One-Hot Encoded Genome Data
Preprocessed Data
Classical Encoder CNN
Extract Latent Features
Quantum Latent Compressor QCNN
Apply Quantum Circuit Operations
Output Quantum Features
Classical Decoder Deconvolution
Reconstructed Genome Sequence

Comparission with CNN /RNN/QCNN



Metric	CNN	QCNN	RNN (LSTM)	GenCoder
Accuracy	75.5%	86.7%	78.84%	86.9%
Training Time (10 Epochs)	4.4 min	4.1 min	5.5 min	7.5 min
Test Loss (MSE)	0.021	0.015	0.019	0.013
Min Test Loss (Per Batch)	0.018	0.012	0.016	0.011
Max Test Loss (Per Batch)	0.027	0.020	0.023	0.017
Evaluation Time	0.9 sec	1.8 sec	1.5 sec	1.4 sec
Trainable Parameters	2.1M	2.3M	3.4M	3.1M

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
1	DeepCGP: A Deep Learning Method to Compress Genome-Wide Polymorphisms for Predicting Phenotype of Rice DOI: https://doi.org/10.1016/j.eswa.2023.119641	Tanzila Islam , Chyon Hae Kim , Hiroyoshi Iwata , Hiroyuki Shimono , and Akio Kimura JUNE 2023	▪ BASIC AUTO ENCODER IS USED ▪ CGP(Cartesian Genetic Programming) ▪ Compression Modeling ▪ Random Forests (RF) ▪ GBLUP and BayesB	▪ Introduces a novel approach for predicting viral genomes in phenotype of rice. ▪ Uses various algorithms and methods for compressing along with ENCODER	Nothing much demerits have been found in this paper	the need for further exploration and comparison of the proposed approach with existing methods for viral genome prediction. Additionally, the paper could benefit from discussing the limitations or challenges encountered during the implementation of the proposed methodology

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
1	Viral genome prediction from raw human DNA sequence samples by combining natural language processing and machine learning techniques DOI: https://doi.org/10.1016/j.eswa.2023.119641	Mohammad H. Alshayeji , Silpa ChandraBhasi Sindhu, Sa'ed Abed journal homepage: www.elsevier.com/locate/eswa 28 January 2023	• The study utilized traditional ML classifiers such as extreme gradient boosting (XGBoost), K-nearest neighbors (KNN), and support vector machine (SVM) to classify and predict viral genomes in DNA sequences. • It employed k-mer counting and the bag-of-words technique to process and analyze the DNA sequences, breaking them down into manageable components for further analysis.	• Introduces a novel approach for predicting viral genomes in human DNA using a combination of NLP and ML techniques. • Presents a model that effectively identifies viral genomes in DNA sequences, demonstrating high accuracy and potential for early diagnosis and treatment of viral illnesses.	**Should work on more varied virus families to test the proposed method's credibilty.**	the need for further exploration and comparison of the proposed approach with existing methods for viral genome prediction. Additionally, the paper could benefit from discussing the limitations or challenges encountered during the implementation of the proposed methodology

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
2	DNA Sequence Classification with Compressors DOI: https://doi.org/10.48550/arXiv.2401.14025	Sukru Ozan January 26 2024	• Compression algorithms like LZMA, Brotli, Gzip are used. • K-NN model is used for enhancement	• Finding the best compression algorithm using accuracy,F1 score,recall, precision and computation time	• The paper only used 7 major compression algorithms:Gzip,Brotli,LZMA,LZ2,BZ2,ZStandard, and Snappy	• The paper finds a very efficient way to compress the genome sequence without any data loss by comparing the original sequence with compressed.

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
3	DeepCGP: A Deep Learning Method to Compress Genome-Wide Polymorphisms for Predicting Phenotype of Rice DOI: 10.1109/TCBB.2022.3231466	Tanzila Islam , Chyon Hae Kim , Hiroyoshi Iwata , Hiroyuki Shimono , and Akio Kimura Published on June 2023	•Compression: Uses deep autoencoders to compress genome-wide polymorphism data. •Prediction: Predicts phenotypes using Random Forest (RF), GBLUP, and BayesB. •Datasets: Two rice datasets (C7AIR, HDRA) with SNP and trait data. •Implementation: Keras and TensorFlow; high-performance computational systems.	•Achieves up to 98% compression with minimal accuracy loss. •Supports large datasets efficiently. •Open-source and replicable. •Retains high prediction accuracy	•Time-intensive training for large datasets. •Limited to rice datasets. •Computational cost of BayesB for uncompressed data. •Information loss at extreme compression.	•Test on other species like humans or other crops. •Add SNP selection capabilities. •Integrate environmental variables. •Reduce training time.

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
4	SQUEEZE AND LEARN: COMPRESSING LONG SEQUENCES WITH FOURIER TRANSFORMERS FOR GENE EXPRESSION PREDICTION Link: https://gattanasio.cc/publication/2023-squeeze-and-learn/	Vittorio Pipoli Giuseppe Attanasio Marta Lovino Elisa Ficarra Accepted on 23 August 2022	▪ Sequence Embedding with Convolutional Layers ▪ DFT-Based Compression ▪ Transformer with Multi-Head Attention (MHA):	•Enhanced Computational Efficiency •Superior Compression with Minimal Information Loss •Improved Prediction Performance	•Dependency on Specialized Hardware for Training •Potential Loss of High-Frequency Information •Limited Generalization to Diverse Sequence Types	• Limited exploration of alternative compression techniques beyond Fourier transforms. • Lack of extensive testing on diverse genomic datasets. • Minimal analysis of the model's robustness to noisy or incomplete DNA sequences. • Insufficient focus on real-time applications and scalability for large-scale genomic studies.

CONCLUSION

Through this research work we have found that QCNN performs well than normal CNN and RNN, Eventhough the implementation is harder ther results are better in both computation wise as well as storage wise