Mental Health Prediction using Quantum-enhanced Machine Learning

Neha Shaikh, Praveen Kumar Thanniru, Sadakhya Narnur, and Rahul Reddy Parupati

Abstract—Mental disorders are more prevalent than heart diseases and diabetes. But there is very little importance given to this issue. With the advancements in AI, this could be an area that will benefit people to create awareness about their mental health. Combining the predictive ability of machine learning with knowledge of which factors could possibly affect mental well-being, a prediction could be made. Though a number of prior researches were made in this field, due to the dimensional constraints very few features were considered for prediction. This reduction in features can affect the accuracy of predictions which will be addressed in this project. The key is to make predictions in a higher dimensional feature space and analyze how efficient and accurate they are. Quantum enhanced Machine learning algorithms are employed to quantitatively estimate mental health based on a number of factors like age, sex, working condition, job, coworkers, etc. A mental health dataset with 27 features is used and modeled with QSVM (Quantum Support Vector Machine). Qiskit, an open-source SDK package in python is used for working with these algorithms. An assessment of the prediction accuracy and precision is made to gauge the performance of QML models chosen for this problem.

Index Terms—Quantum-enhanced ML, QSVM, Curse of Dimensionality, Hilbert Space

1 Introduction

The growing stress and responsibilities are often presumed to be common in our busy lives. However, to what extent it affects our Mental health is often overlooked. Many don't realize the risk of them being on the edge of mental illness. We often come across biological mental disorders that have genetic factors involved. But mental health is a broader spectrum that involves any psychological imbalance that starts to affect our thoughts, feelings, and actions. Despite one never facing adverse life experiences like trauma, abuse or violence can still be vulnerable. This makes it clear that a person's mental health changes with time, based on many factors. Some of the factors can be early life experiences, strenuous working hours, economic burdens, social anxiety, and even physical health. This leads to a need for awareness of emotional and mental well-being. An assessment of one's sanity from time to time would help in alerting them of any need for help or treatment and thereby directing their lifestyle for overall health.

This project is based on idea where a prediction could be made of mental health as a consequence of multiple factors like age, family history, working conditions, employment, etc. Putting Machine learning into work for this project would help in predicting with greater accuracy. In ML predictions a larger number of features increase predictive modeling tasks. In order to overcome this computational overhead, dimensionality reduction referring to the reduced number of input variables is often considered. However, dimensionality reduction could be seen as a drawback that affects the predictions, often called the Curse of Dimensionality. Hence in order to extend the prediction from the euclidean vector space to higher dimensional Hilbert space to consider more features, Quantum computing will be introduced to machine learning which is Quantum-enhanced Machine learning. This can be seen as the execution of machine learning computations on a quantum computer. It can be seen as an enhancement over machine learning prediction as it not only overcomes the dimensionality limitations but also does the computations with greater speed and improves data storage done by algorithms.

1.1 Objective

The objective is to prove that the QEML method will enhance and improve the ability of the traditional machine learning algorithms. The data set is selected from the CDC website that describes different features which can be considered risk factors in mental health predictions. Aim would be to explore all the possible correlations between the features and use as many required for better prediction of whether treatment is required for a particular person's record.

To address the problem statement, we are using a super powerful quantum variants of machine learning classification algorithms: Support Vector Machine through hamming distance metrics in quantum feature space.

- *Neha Shaikh was with the Department of Applied Data Science, San Jose State University, San Jose, CA, 95192. E-mail: neha.shaikh@sjsu.edu*
- *Praveen Kumar was with the Department of Applied Data Science, San Jose State University, San Jose, CA, 95192. E-mail: praveenkumar.thanniru@sjsu.edu*
- *Sadakhya Narnur are with the Department of Applied Data Science, San Jose State University, San Jose, CA, 95192. E-mail: sadakhya.narnur@sjsu.edu*
- *Rahul Reddy Parupati was with the Department of Applied Data Science, San Jose State University, San Jose, CA, 95192. E-mail: rahulreddy.parupati@sjsu.edu*

Manuscript received November 28, 2022

2 Theoretical bases and literature review

2.1 Definition of the problem

This study focuses on predicting the mental health of a person based on the features. In our study, we will treat the classic Machine Learning algorithms like Support Vector Machines with Quantum Computing and analyze the degree of risk a person experiences regarding his/her mental health issues.

2.2 Theoretical background of the problem

Pattern recognition is used in machine learning, which is a data-dependent approach for datasets of a certain size. However, as the size and quantity of features for the given case study increase, performance reduces clearly for both regression and classification machine learning algorithms. This issue is also referred to as the Curse of Dimensionality, has severely increased runtimes and caused quadratic growth in machine learning methods. This issue stems from how data is stored in the first place. The states and characteristics of specific vectors are positioned in classical feature space for classical machine learning. The kernel operations and performance of some machine learning algorithms are inhibited by exponential storage and runtime due to this feature space. To ease the strain of exponential storage and so create a quantum feature space, it has recently been suggested to use quantum computing. The simplest answer is to convert classical states into quantum states, which can be stored more effectively.

2.3 Literature Review

The current problem falls under the classification and in the presence of a number of algorithms for classification it is considered that SVM and KNN besides Random Forest are the best performing. This study tried clustering models to identify the number of possible clusters before deciding on modeling[1].This research was more focused on classifying different target groups that were vulnerable.

In a similar research Predicting Mental Health Illness [2] using Machine Learning algorithms it was carried out with five algorithms namely Logistic Regression, K-NN Classifier, Decision Tree Classifier, Random Forest and Stacking. The study shows k-NN had comparably good accuracy for classification.

In the research [3] a brief introduction on Quantum enhanced machine learning is given as a classic machine learning in the high dimension space. QSVM is implemented on breast cancer dataset which is a binary classification problem. It is observed that Quantum enhanced SVM performed better than classical ML SVM.A possible limitation on QSVM is discussed regarding the implementation of kernel models as it would be complex to implement using Quantum circuits for Radial basis kernel.

A paper on Investigation of Quantum Support Vector Machine for classification [4] made a comparision of Quantum SVM performance over a quantum computer. They have encoded the data and run on QSVM circuit and later checked for the performance.They have observed that the QSVM has not performed well when they implemented it on NISQ era quantum simulator. They got a mere accuracy of 62%. However later they proposed a mechanism which gave an improvement over it.

2.4 Solution for our problem

In this study, we as a team studied the Support Vector Machines methodology and understood the functionality of quantum-enhanced feature space. Following that, we declared the fundamental characteristics of quantum machine learning as a whole and the justification behind why it should outperform the conventional SVM algorithm. Thus the prediction of risk factors involved in the mental health of employees can be carried out effectively with the involvement of Quantum Computing.

2.5 Why our solution is better

With the application of a quantum feature space, which intensifies current Machine Learning algorithms as it uses the parallelization and the reduction of the storage space from exponential to linear, we may be able to solve the problem of determining the risk factor by incorporating quantum circuits into conventional Machine Learning, instead of using the traditional approach alone.

3 Methodology

3.1 Collecting the input data

The data in focus is any mental health related data that supports the classification. The Mental health dataset that is used for the project is taken from a survey conducted by Chirag Dodiya [5]. The survey was conducted where the responses of the people were recorded for classification purpose.

3.2 Exploratory data analysis

Understanding the data we are going to work with helps build an efficient model. This is the second step in our project. Post understanding the project scope and defining the problem statement we have decided to use the mental health dataset for classifying a person's record. The dataset has 1260 records over 27 features namely: Timestamp, age, gender, country, state, self-employed, family history, etc.The first step of the analysis was to find the missing values in the dataset. A column summarizing the count of missing values has been observed like in Fig 1.

State and comments are the two columns with the highest missing values. Further work _interfere and self_employed had some missing values. The gender column has a number of entries that need to be normalized and reduced to some subgroups. The age column had 4 outliers with values out of the range of the acceptable age.

A correlation matrix heatmap is used to visualize the correlations between all the features in Fig. 2. A strongest correlation is observed for treatment and work_interference which aided us to focus more on that feature. Family_history, benefits and care options are the next positive correlated. A distribution of age is observed with most of distribution having been teenagers. Supporting this a bar chart plotted for the age groups requiring treatment shows

Timestamp	0
Age	0
Gender	0
Country	0
state	515
self_employed	18
family_history	0
treatment	0
work_interfere	264
no_employees	0
remote_work	0
tech_company	0
benefits	0
care_options	0
wellness_program	0
seek_help	0
anonymity	0
leave	0
mental_health_consequence	0
phys_health_consequence	0
coworkers	0
supervisor	0
mental_health_interview	0
phys_health_interview	0
mental_vs_physical	0
obs_consequence	0
comments	1095

Fig. 1. Dataframe showing the null values count for different columns

that people with ages 7 to 20 are showing highest need for treatment.

The distribution of ages and their count can be seen in Fig. 2 visualized as a histogram. The density or a larger portion of records are observed for age groups 10 - 20.

Nested barplots visualizing probability of mental health condition clearly show that in ages 0-20 Trans gender, gender queer in 21-30 and female in old ages have most probability.

3.3 Data cleaning

Dropping columns - Some of the features like Timestamp, state and comments do not contribute towards the classification. Furthermore, state and comments have large missing values. Hence it would be logical to drop these columns when classifying the data.

Handling missing values: On dropping the two main missing value columns we were left with self_employed and work_interfere to handle the missing values. This is handled by assigning 'No' to the self_employed column in missing places based on our exploration of records. For handling work_interfere we added a new entry named Don't Know for records with missing values.

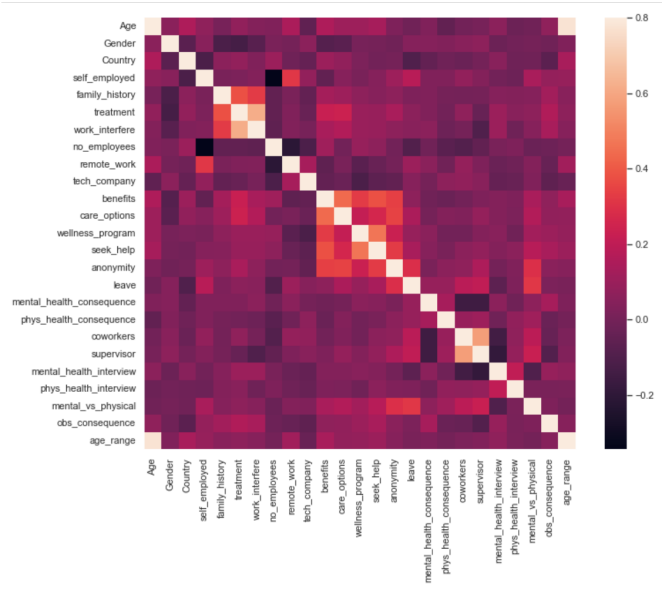

Fig. 2. Heat Map showing feature correlations

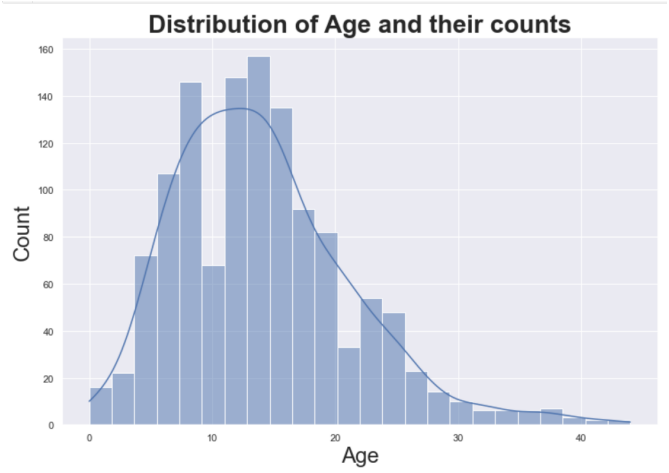

Fig. 3. Histogram showing distribution of persons over age

Handling outliers: The age column has 4 outliers which are handled by replacing the outliers with the column's median age.

The gender column had different entries which had spelling mistakes, short forms, synonymous words, etc. We had to normalize all the entries into a few sub-groups like Male, Female, Gender queer and Trans.

3.4 Data Pre-Processing

A number of the features considered are categorical hence they are required to be encoded into numericals. Label encoder is used to encode the Treatment feature which is the y in case of the classification and min max scalers and binarizers are used for encoding the features.

4 Problem Solution

4.1 Algorithm design

A number of Machine learning algorithms could be used for this classification provided the number of features are

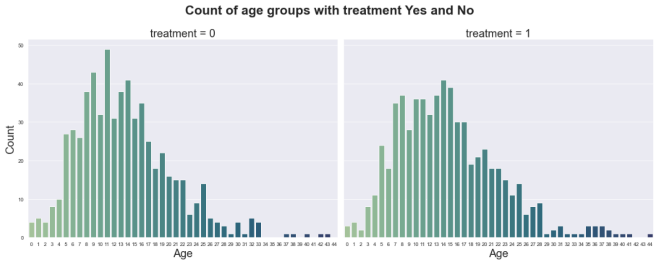

Fig. 4. Plot showing the count of age groups who require treatment and not

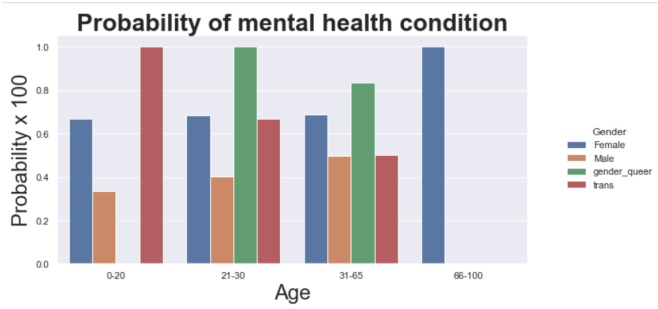

Fig. 5. Probability of mental health conditions over age

limited to avoid the curse of dimensionality. This limitation results in the reduction of possible features which could contribute in a way or other to the classification. This could be handles using quantum computers which are expensive and complex to run. The approach we have considered is Quantum enhanced ML to give the power of Quantum computers yet run them in a simple way.

4.1.1 *Quantum Support Vector Machines*

As we saw in traditional the advantage of using traditional SVM is to classify data points in multiple dimensions by moving from euclidean space to Hilbert space. However, it has certain limitations on the number of dimensions it can handle using the kernel trick, and optimizing the kernel trick is very difficult with traditional SVM. Here is where QSVM helps with keeping the data points in the quantum state which can be achieved by a swapping circuit and building the kernel of SVM from these quantum states. After calculating the Kernel matrix on the quantum computer they can train the Quantum SVM the same way as a classical SVM.

The QSVM algorithm overview can be seen in Fig. 7 where three qubits where the first one is training register, second is input —y¿ state and third is ancilla qubit.

4.2 Modeling

4.2.1 *Quantum SVM*

The Qiskit SDK package by IBMQ simulation products was used to implement the quantum variant of the traditional SVM. The documentation on the website is what was referred to while modeling the dataset. The QSVM approach is used for classification problems that call for a feature map but for which standard kernel computation is inefficient. As a result, it is anticipated that the needed computer resources

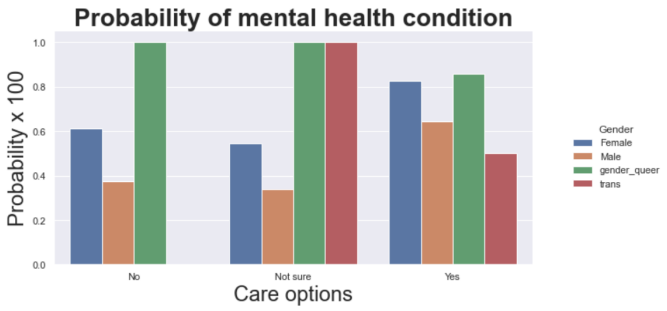

Fig. 6. Probability of mental health conditions over care options

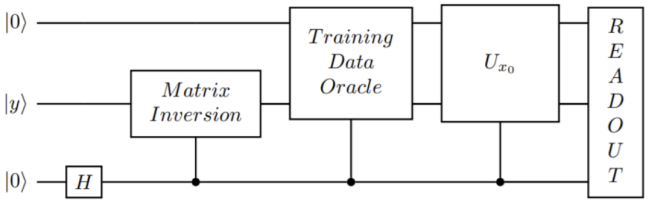

Fig. 7. QSVM algorithm overview with readout

will grow exponentially in proportion to the problem's size. In order to tackle this issue, QSVM directly estimates the kernel in the feature space using a quantum processor. The technique is characterized as "supervised learning," and it consists of a training phase (when the kernel is calculated and the support vectors are obtained) and a test or classification phase (where new data without labels is classified according to the solution found in the training phase). Depending on how many classes the data has, QSVM will internally perform a binary classification or a multiclass classification. If the data contains more than two classes, a multiclass extension must be provided. Below are the steps performed while implementing the algorithm. 1) Perform train, and test split on our dataset. 2) Plot the feature maps. 3) assign a feature map, and calculate the fidelity. 4) Select an SVM kernel. 5) fit and test the QSVM.

Algorithm 1 QSVM Pseudo code for featuremap selection

Input: Select training samples $(\vec{}_i, y_i) : \vec{}_i \in R^N, y_i = \pm 1_{i=1,2,3...n}$
Output: Best fit feature map selection

Parameters: Train and test sample, featuremap, repetitions(reps), entanglement, multiclass
while $reps = 1 \Leftarrow n$ **do**
  for $Pauli\_list = 1 \Leftarrow 10$ **do**
    $U_\Phi(\vec{x}) = exp(i \sum_{j=1}^{n} \alpha_j \phi_s(\vec{x}) \Pi \sigma_j \in \{I, X, Y, Z\})$
      Record measurement through ancilla register
  end for
end while

Fig. 8. First algorithm in quantum variational support vector machine

4.3 Evaluation metrics

Our classification performance was planned to be measured using a number of metrics namely accuracy, precision, recall, and F1 score. However the qiskit library we used had not many metrics and we had to settle with the accuracy score alone. Lets look at the metrics which could be used for

evaluation. The general terms for any prediction would be True Positive (TP), True Negatives (TN), False Positive (FP) and False Negative (FN) Conditional Positive = P = TP + FN Conditional Negative = N = FP + TN

4.3.1 Accuracy

Accuracy can be seen as the closeness to the true values of the measured quantity. It mainly measures the statistical bias. It considers both the positive and negative samples being classified, thereby measuring the ability to classify positive samples in model.

Accuracy = (TN + TP) / (TN+TP+FN+FP) = (TN + TP) / (P + N)

4.3.2 Precision

Precision is measure of reproducibility of the measurement. It is a measure of statistical variability.

Precision = (TP)/(TP+FP)

4.3.3 Recall

Recall identifies accurately classifying all positive samples while ignoring if any negative samples classified as positive. It considers only the positive classifications while ignoring all negative samples. This allows us to identify how many positive classifications are made.

Recall = (TP)/(TP+FN)

4.3.4 F1 score

This is a combined measure of precision and recall of a classification model as a harmonic mean. This is the best way to compare two models when a decision cannot be made based on the recall and precision.

F1 = 2*(precision * recall)/(precision + recall)

4.4 Model comparision

QSVM has shown a pretty good performance with a score of 0.9 for the mental health dataset when 10 features where considered. On the other hand conventional SVM gave a training accuracy of 0.72 but a very low testing accuracy of 0.50 for the same 10 features.

4.5 Languages Used

The project is developed in Python Programming Language. Core modeling of algorithms SVM is implemented in Python.

4.6 Tools used

Google colab notebook is used for executing the models. Sklearn is used for preprocessing of data, matplotlib and Seaborn are used for visualizations, Qiskit is used for Quantum powering the models and sklearn models for comparision of performance.

5 Conclusion

The conclusions should only be regarded with a grain of salt because these results were computed on a small dataset with n = 10 dimensions and were somewhat random. When quantum computing becomes more widely available, we might need to perform a cost analysis to determine whether the advantage is worthwhile or not. Physical quantum computers are now only available to academics and collaborators at accredited institutions, not to the general public. As a result of the notion of an improved quantum feature space, machine learning can thus benefit from quantum computing.

Recognizing that modern quantum computers are highly vulnerable to noise and physical perturbations during calculations is another crucial concern. Decoherence is a significant issue as well since it will not be practicable if physical quantum computing is ever made available to the majority of researchers. Today's quantum computers frequently lose or collapse quantum information.

Appendix A
Criteria met in Rubrics

1. Visualizations - During the exploratory data analysis phase a number of visualizations are used to understand the data and show correlations.

2. Presentation skills - We have practiced to maintain and cover important processes and concepts involved in our project by allotting enough time for questions.

3.Significance to the real world - This project is towards a social cause to bring awareness and in a way help people to understand their mental health condition The outcome of the project will inspire a similar idea of using QEML for other social causes and will improve the performance of traditional ML algorithms with the help of QEML.

4. Saving the model for quick demo - Our models have been saved and can be loaded for testing and demo purpose. We have used picking approach where the trained model is saved as a file and later which can be loaded and test data can be run.

5. Code Walkthrough - The complete code is available in the colab notebook with clear comments making it easy to understand and elaborating the operations performed.

6. Report - We have strictly adhered to IEEE format for our report while taking care of the language and keeping it our own work. The report includes all the details required to understand our problem and approach towards solving it.

7. Version Control - Our entire data and code is maintained in Github repository with readme file.

8. Discussion / QA - Throughout the presentation, the discussion would be open and questions would be answered. There will be some time allocated for QA session in the end of the demo.

9. Lessons learned - The project is selected to be an added knowledge to our learning. With the understanding from the class curriculum about SVM we have extended thier applications in Quantum enhanced ML. Literature review, background, conclusions and comparisions all of the sections include the learnings from the project.

10. Prospects of winning competition / publication - QEML is a topic still under research and needs a lot of researchers to contribute. This could be a potential paper for publication as a very few references are available on QEML based models and especially for our classification problem. With some improvements in our approach this can be in future capable of a competition.

11. Innovation - QEML is a very new concept that needs to come into light. Our project is an implementation of this rarely explored concept to a social cause. We attempted to implement few algorithms that weren't previously available which could be seen as an area of future research.

12. Evaluation of Performance - The models are checked on an evaluation metric. Initially the correlations were checked which was a partial evaluation and later a comparision is made between models performance.

13. Teamwork - The whole team was involved in each phase of the project and had a weekly sync up to discuss progress towards the goal. A clear credit statement is mentioned in Appendix B.

14. Technical difficulty - As the topic is pretty new, less number of research papers were available to refer to and not many have addressed the problem or use case, also this requires the implementation of complex python libraries which simulate Quantum Computing power. A very few traditional machine-learning algorithms are implemented using this QEML method hence the modeling phase took a reasonably longer time. We have tried to implement few other models which weren't available in the library which was pretty difficult and had to drop them.

15. Practiced pair programming - GitHub copilot is used as a plugin in VScode.

16. Practiced agile / scrum (1-week sprints) - We followed a 1 week sprints agile framework using JIRA. Weekly meetings were held to track our progress on the deliverables.

17. Used Grammarly for language - We have used Grammarly to check our document language and rules.

18. Slides - Prepared our presentation slides to cover important aspects of the project.

19. Demo - We have prepared a demo structure to follow for a better presentation and show the working model.

20. Using LaTEX - We have used laTex for formatting our report using the IEEE template. All the documentation is done using overleaf.

21. Used creative presentation techniques - The presentation slides are prepared with the Prezi tool with interesting animations.

22. Literature Survey - We have referred to papers and research that cover mental health predictions and classifications. All the literature is distributed into meaningful subsections and concepts are introduced appropriately. Al the cited works are referred.

APPENDIX B
AUTHOR CONTRIBUTIONS

Sadakhya Narnur: Conceptualization, Methodology, Investigation, Writing – Original Draft. Neha Shaikh: Software, Validation, Data Curation, Writing – Review and Editing. Rahul Reddy Parupati: Visualization, Formal analysis, Validation, Writing – Review and Editing. Praveen Kumar Thanniru : Project administration, Investigation, Writing – Review and Editing.

ACKNOWLEDGMENTS

The authors would like to thank Professor Dr. Vishnu Pendyala, Nikita Balani and Vineeth Reddy Chintala for being supportive throughout the process. This work is done as a part of Machine Learning Technologies course work.

REFERENCES

[1] Sumathi, M. R., and B. Poorna. "Prediction of mental health problems among children using machine learning techniques." International Journal of Advanced Computer Science and Applications 7.1 (2016).

[2] Vaishnavi, Konda, et al. "Predicting Mental Health Illness using Machine Learning Algorithms." Journal of Physics: Conference Series. Vol. 2161. No. 1. IOP Publishing, 2022.

[3] Thomas, Dr G., Krishna Sai Mangalarapu, Munawar Ali Md, and Vamsi Krishna Talakokkula. "A New Quantum Approach to Binary Classification." arXiv preprint arXiv:2106.15572 (2021).

[4] Kariya, Anekait, and Bikash K. Behera. "Investigation of Quantum Support Vector Machine for Classification in NISQ era." arXiv preprint arXiv:2112.06912 (2021).

[5] Chirag Dodiya.2021. Mental-Health-Prediction-using-Machine-Learning-Algorithms. https://github.com/cdodiya/Mental-Health-Prediction-using-Machine-Learning-Algorithms/blob/main/survey.csv(2022)

Neha Shaikh is currently pursuing a Master of Science in Data Analytics. The main research areas include sentiment analysis, enhancing ML algorithms, and drug discovery. Passionate about using ML for social good.

Praveen Kumar is currently pursuing a Master of Science in Data Analytics.

Sadakhya Narnur is currently pursuing a Master of Science in Data Analytics. Her main research interests include Object Detection and prediction & analytics in the field of medicine.

Rahul Reddy Parupati is currently pursuing a Master of Science in Data Analytics.