Quantum-Enhanced Data Analytics For Crime Prediction

Abraham Ighalo
School of Computer Science and Technology
Algoma University
Saulte Ste. Marie, ON, Canada
aighalo@algomau.ca

Ajmery Sultana
School of Computer Science and Technology
Algoma University
Brampton, ON, Canada
ajmery.sultana@algomau.ca

***Abstract*—As criminal tactics become increasingly sophisticated, using advanced tools and technologies, the demand for more powerful analytical systems has surged. Although classical machine learning approaches have been widely used in criminal analysis, the exponential growth of data, driven largely by large language models (LLMs) and AI agents, has exposed their limitations. Quantum computing, rooted in the principles of quantum mechanics, presents a promising alternative that harnesses the behavior of atoms and electrons to perform complex computations. This study evaluates classical and quantum machine learning approaches for crime prediction, analyzing their efficacy in classifying criminal incidents. Classical models, such as logistic regression, XGBoost, random forest, and support vector machine, demonstrated robust performance, with ensemble methods such as random forest and XGBoost achieving particularly high effectiveness in classification tasks. In contrast, quantum models, including the variational quantum classifier, the quantum neural network, and the quantum support vector machine, showed promising theoretical advantages but lagged in practical performance due to current hardware. The findings highlight the need for further refinement of quantum techniques to bridge this gap and fully harness their potential for complex pattern recognition in criminal data analytics.**

***Index Terms*—Quantum machine Learning, Quantum Data Analytics, Predictive Policing, Crime Detection, Classical machine Learning.**

I. Introduction

Crimes have been around since the early days of human development; in the time of the Babylonians, the cost of committing a crime was met with retribution, “an eye for an eye”, a tooth for a tooth [1]. In modern times, terms such as the criminal code have been used to inform and guide citizens of any country on what is lawful and what is not. Criminal activities destabilize growth and break the unity of a country, as they introduce fear in the citizens and hinder them from freely enjoying their lives. In Canada, the rise of fraud in this has skyrocketed since 2012. In 2022, Statistics Canada recorded that fraud cases in the country had doubled in the past ten years, from 79,000 in 2012 to 150,000 in 2022. [2]. The reputation of countries are also affected by high rates of criminal activity. Countries with high rates of criminalism tend to be labeled as “fraudulent countries” and conceived as unsafe and unaccommodating to live in. On a global stage, countries are ranked by the IEP (Institute of Economics and Peace) on the basis of their crime rates. This raking could damage any country’s reputation, so it is important that government leaders and law enforcement agencies ensure that criminal incidents are kept in check.

In recent years, machine learning (ML) has emerged as a powerful tool for crime prediction and prevention, leveraging historical data to identify patterns and forecast criminal activity. Techniques such as logistic regression, random forests, and support vector machines (SVMs) have been used to classify types of crime, predict hotspots, and optimize resource allocation for law enforcement [3]. These models analyze structured data, such as incident reports, geographic coordinates, and temporal trends, to generate actionable insights. For example, ensemble methods such as XGBoost have demonstrated high accuracy in classifying crime severity, while computer vision algorithms automate surveillance analysis, reducing manual review burdens [4]. However, classical ML approaches face limitations in handling exponentially growing datasets and complex, high-dimensional feature interactions, prompting exploration of more advanced computational paradigms.

Quantum computing, with its inherent parallelism and ability to process information in superposition, offers transformative potential for data analytics. By exploiting quantum properties such as entanglement and interference, quantum algorithms can theoretically solve certain problems, such as optimization and pattern recognition, exponentially faster than classical systems [5]. In crime prediction, quantum machine learning (QML) models such as variational quantum classifiers (VQCs), quantum neural networks (QNNs), and quantum support vector machine (QSVM), can uncover latent patterns in large-scale datasets while reducing computational overhead. Despite these advantages, current quantum systems remain constrained by hardware limitations [6]. This paper bridges the gap between theory and practice by empirically evaluating different classical and quantum ML models on the New York City Police Department (NYPD) Complaint Dataset, assessing their efficacy in crime classification, and identifying pathways for future optimization.

II. Related Work

The rapid evolution of QML has generated significant research interest in recent years. This section critically evaluates key contributions to the field, analyzing their methodologies, findings, and implications for practical applications. The study in [6] introduces a Quantum K-Nearest Neighbors (QKNN) classifier, leveraging quantum superposition and entanglement to outperform classical KNN in the Wisconsin Breast Cancer Dataset (10 features). Using the Hamming distance metric and Grover's algorithm in a simulated IBM quantum environment, QKNN achieved 97.18% accuracy vs. KNN's 96.27%. Key advantages include reduced feature space via qubit superposition and exponential speedup through quantum parallelism. Limitations include decoherence, hardware constraints, and a small dataset size. This demonstrates the potential of quantum computing to improve machine learning tasks such as binary classification.

A Quantum support vector machine (QSVM) is applied in [7] to classify 2.4 million financial transactions, comparing it with classical models (XGBoost, random forest). Despite using only 7 features (vs. 37 for classical methods), QSVM achieved a 1% accuracy boost, highlighting the quantum potential in fraud detection. Data preprocessing included rule-based models, classical ML, and quantum analysis, with undersampling for class balance. The results suggest that quantum performance could scale with expanded feature access, though current hardware constraints limit practical deployment.

The studies in [8]–[10] present a Quantum Convolutional Neural Network (QC-CNN) that combines quantum circuits with classical CNNs to analyze surveillance footage for criminal anomalies such as robbery and vandalism. Using a dataset of 1,900 video clips (128 hours, 7,247 frames), the hybrid QC-CNN model demonstrated superior performance compared to classical methods and other quantum-classical approaches (such as Quantum Generative Adversarial Network (QGAN)), achieving the highest scores across accuracy (97.8%), precision (96.5%), recall (98.2%), and F1-score (97.3%) metrics [9], [10]. The quantum layer's ability to process complex patterns in video data addressed key limitations of classical systems, including high false alarm rates and scalability challenges [8]. These results highlight the potential of quantum-enhanced neural networks for real-time crime detection in law enforcement applications, supporting a broader investigation of quantum advantages in security analytics.

III. Dual-Path Methodology

This section describes a dual-path framework to empirically evaluate different classical ML (logistic regression, XGBoost, random forest, SVM) and QML models (VQC, QNN, QSVM) for crime classification using the NYPD Complaint Dataset. The methodology systematically compares their performance through data preprocessing, feature engineering, and cross-validation while addressing quantum-specific constraints through hybrid optimization techniques. The workflow for this dual-path methodology is illustrated in Fig. 1.

A. Dataset

In this work, the NYPD Crime Complaints dataset is analyzed. The NYPD Complaint Dataset is a historical dataset that contain information from 2006 until April 23rd 2024. This data includes all crimes reported to the NYPD. The information documented in this dataset includes key details of the type of crime committed, information about victims, as well as the key features of suspected criminals. The NYPD made this resource public in 2016 and has been maintaining the data since. At the time of this paper, the total number of rows is 8.91M, with 35 columns, each row representing a complaint made by a resident.

B. Data Preprocessing

To optimize model performance, the NYPD Complaint Dataset underwent rigorous pre-processing to address data quality challenges. Key steps included null-value imputation, feature engineering, and dimensionality reduction, critical for both classical machine learning and quantum-ready data encoding.

- Used Dask's lazy loading to handle 8.91M rows efficiently.
- Duplicate copies were removed, missing values were handled, and the formats were standardized.
- Investigated and addressed outliers.
- Identified and transformed skewed columns using mean, median, square root, cube root, and Yeo-Johnson techniques.
- Applied imputation based on skewness results (mean for symmetric, median for skewed).
- Conducted correlation analysis using heatmaps.
- New features developed from existing columns.
- Class balancing using the random OverSampler class from Scikit-learn
- Encoded categorical and numerical columns using:
  - Label Encoding for ordinal data
  - One-Hot Encoding for categorical data

IV. Parallel Processing Paths

A. Classical ML Models

This section details the experimental framework for evaluating classical ML models in crime classification tasks. Our primary research objective is to assess whether precinct-level crime features can effectively predict legal severity categories (felony, misdemeanor, or violation) without relying on internal crime description metadata. This study employs supervised classification to predict the severity categories (felony, misdemeanor, or violation) of the crime using the NYPD dataset. The research addresses four key questions:

- Predictive capability of precinct-level features without crime descriptions
- Officer allocation efficiency through crime completion patterns
- Victim profiling by demographic characteristics
- Temporal and spatial crime forecasting (2026 safety projections)

As a labeled-data problem with discrete outcomes, we implement classifier models rather than regressors. logistic regression serves as our baseline algorithm, processing crime

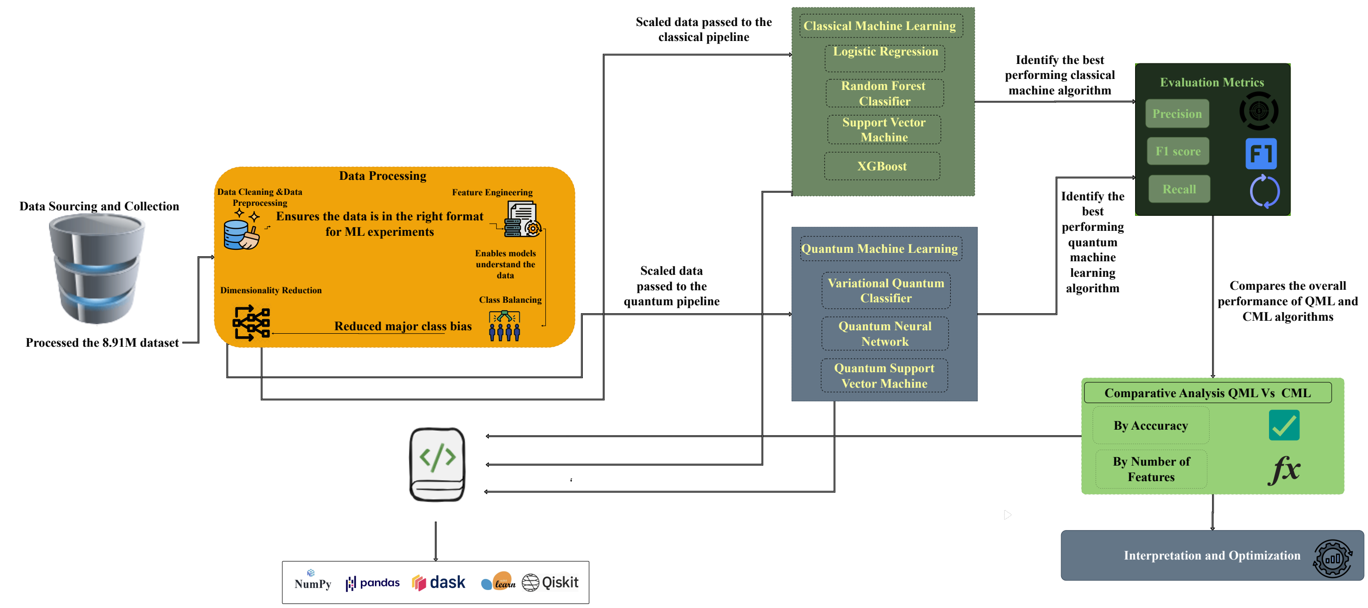

Fig. 1: Workflow for the dual-path methodology

location and completion status features to establish classification boundaries between legal severity categories [11]. This discrete value approach contrasts with continuous regression tasks [12], aligning with our multiclass prediction objective.

The variables selected from the NYPD dataset reflect key dimensions of reported incidents, including contextual, demographic, and temporal aspects relevant to the analysis. The following are the independent variables used in the work, categorized by type:

Independent Variables: independent variables used in the work, categorized by type:

- **Crime and Location Information:**
  - `CRM_ATPT_CPTD_CD` – Crime attempted or completed
  - `ADDR_PCT_CD` – Address precinct code
  - `LOC_OF_OCCUR_DESC` – Location of occurrence description
  - `BORO_NM` – Borough name
  - `PREM_TYP_DESC` – Premise type description
  - `Y_COORD_CD` – Y coordinate (used for mapping)
  - `Longitude` – Longitude coordinate
- **Suspect Information:**
  - `SUSP_RACE` – Suspect race
  - `SUSP_SEX` – Suspect sex
  - `SUSP_AGE_GROUP` – Suspect age group
- **Time Information:**
  - `Month_Abbr` – Month abbreviation
  - `CMPLNT_FR_DT` – Complaint date

Dependent Variable:

- `LAW_CAT_CD` – Legal category of the crime (e.g., felony, misdemeanor, violation)

Classical ML Models

- **Logistic Regression Model**: The Logistic regression model is one of the most common base linear models used in ML. This ML algorithm predicts the statistical probability of an outcome based on a set of independent variables. It is useful in classification and predictive tasks to predict the binary outcomes or labels of an event, hence 0's and 1's, guilty not guilt. Multinomial logistic regression is a classification method which generalizes the logistic regression from its binary classification o's and 1's to more discrete outcomes or possibilities, hence it is able to classify more than binary classification problems. It uses this generalization of the logistic function also known as softmax and normalizes the data into probability distributions consisting of K probabilities that are proportional to the number of inputs, thus the larger the input, the more possibilities.

$$P(y = c \mid \mathbf{x}) = \frac{e^{\mathbf{w}_c^T \mathbf{x}}}{\sum_{k=1}^{C} e^{\mathbf{w}_k^T \mathbf{x}}}$$

- **XGBOOST Model :** Extreme Gradient boosting also known as XGBOOST is a supervised machine learning algorithm designed for efficiency, speed, and high performance. The models is the optimized version of the gradient boosting algorithm and is an ensemble algorithm which refers to the combination of multiple weak or low performing algorithms to form a much optimized version. With the collection of key features such as ensemble learning, gradient decent optimization, and regularization, this model is efficient for our crime classification task.
- **Random Forest Classifier :** The Random forest classifier is another example of an ensemble algorithm. They combine more of the same algorithms for the purpose

of developing a much-optimized version for classifying labels.

- **Support Vector Machine :** In svm, the aim of the algorithms is to determine the best hyperplane. This hyperplane can be seen as a line separating the features in a dimentional space. In binary classifications, it will be a 2D space with a line sperating the labels. In multiclass classifications, the svm essentially breaks the classifications into multiple binary classifications with the goal of gaining distance between each pair of classes; this is called the one-to-one approach.

Performance Analysis of Classical ML Models The results of these techniques in various classical ML algorithms on a large dataset include:

Mean cross-validation accuracy: 0.6471				
Class	**precision**	**recall**	**f1-score**	**support**
0	0.77	0.93	0.84	3089
1	0.54	0.44	0.48	3089
2	0.58	0.57	0.58	3089
Accuracy			0.64	9267
Macro avg	0.63	0.64	0.63	9267
Weighted avg	0.63	0.64	0.63	9267

TABLE I: logistic regression

Mean cross-validation accuracy: 0.9886				
Class				
textbfprecision	**recall**	**f1-score**	**support**	
0	1.00	1.00	1.00	3089
1	0.99	0.99	0.99	3089
2	0.99	0.99	0.99	3089
Accuracy			0.99	9267
Macro avg	0.99	0.99	0.99	9267
Weighted avg	0.99	0.99	0.99	9267

TABLE II: XGBoost

Mean cross-validation accuracy: 0.9345				
Class	**precision**	**recall**	**f1-score**	**support**
0	1.00	1.00	1.00	3089
1	0.92	0.92	0.92	3089
2	0.92	0.93	0.92	3089
Accuracy			0.95	9267
Macro avg	0.95	0.95	0.95	9267
Weighted avg	0.95	0.95	0.95	9267

TABLE III: Random Forest Classifier

Mean cross-validation accuracy: 0.7172				
Class	**precision**	**recall**	**f1-score**	**support**
0	0.80	0.99	0.88	3089
1	0.67	0.49	0.57	3089
2	0.68	0.70	0.69	3089
Accuracy			0.73	9267
Macro avg	0.72	0.73	0.72	9267
Weighted avg	0.72	0.73	0.72	9267

TABLE IV: Support Vector Machine

B. Quantum ML Models

The study of QML can be divided into four sections depending on the computing resources used in the experiment. It could be either a quantum or a classical system computer. The research of QML algorithms follows the following four paths:

- **CC :** This is the most common use of machine learning algorithms. It refers to the use of classical processes together with classical systems or machines to make predictions and develop models. Examples include the use of artificial neural networks to differentiate between the different types of weapon used in a crime scene or the different types of bruises caused by them.
- **CQ :** Classical process processed in a quantum machine. This research area focuses on the use of classical ML algorithms on classical processors to analyze quantum data collected from quantum systems [13]. According to [13], some examples are when a classical neural network model is performed on quantum systems, measured back to classical states, and used to implement regression or classical experiments [13].
- **QC :** Quantum process processed in a classical computer. Here, quantum advantage is utilized to improve the classical processes of analyzing the patterns in datasets. A good example of this will be to utilize quantum neural networks along with quantum kernels to perform image recognition from security cameras in retail stores during investigations [13].
- **QQ :** Quantum data processed in quantum device to process quantum data [13]. Here, quantum utilities help reduce the computational cost of analyzing and understanding complex quantum systems. This is the future of quantum computing, where quantum data is utilized on a quantum machine.

QML approaches to be investigated include:

- **Variational Quantum Classifier :** Variational Quantum Classifier are a type of hybrid quantum machine learning that is useful in machine learning classifications. It works by first encoding the data that are being processed in the quantum state using the feature map. Feature maps are quantum circuits that have as input the dataset being worked on and the output is a quantum state. This state is then passed to the variational quantum circuit to be parameterized. After parameterization, the next step is to optimize using a loss function; this helps to define how well the variational quantum classifier is able to classify the labels.
- **Quantum Neural Network:** Just like the classical neural networks took inspiration from the way the human brain works, quantum neural networks took inspiration as well from classical neural networks. They applied this by using both the CNN approach and parametrized quantum circuits. As a machine learning algorithm, QNN models are trained on data to determine the relationships or patterns in the dataset. The QNN models encode classical data into quantum states which are then processed by the

quantum gates and trained using parametrized weights **(ansatz)**.

- **Quantum support vector machine :** Classical support vector machines as seen previously rely on determining the optimal hyperplane for the number of labels or features in the dimension space. However, support vector machines are dependent on the distribution of the data. When an optimal hyperplane is unavailable, the svm employs the use of kernels to classify the labels. The kernels are transformation features that simplify classification by mapping data points to new dimensional spaces.

Performance Analysis of QML Models Given the computational demands of quantum algorithms, the results presented below are based on experiments carried out using a dataset consisting of 1,127 samples and 6 features, to maintain consistency and manage resource constraints.

VQC Accuracy: 0.56				
Class	**precision**	**recall**	**f1-score**	**support**
0	0.56	0.59	0.58	125
1	0.57	0.54	0.55	124
Accuracy			0.57	249
Macro avg	0.57	0.57	0.57	249
Weighted avg	0.57	0.57	0.57	249

TABLE V: VQC Results

QSVM Accuracy: 0.63				
Class	**precision**	**recall**	**f1-score**	**support**
0	0.66	0.53	0.59	125
1	0.60	0.73	0.66	124
Accuracy			0.63	249
Macro avg	0.63	0.63	0.62	249
Weighted avg	0.63	0.63	0.62	249

TABLE VI: QSVM Results

QNN Accuracy: 0.50				
Class	**Precision**	**Recall**	**F1-score**	**support**
0.0	0.50	1.00	0.67	125
1.0	0.00	0.00	0.00	124
Accuracy			0.50	249
Macro avg	0.25	0.50	0.33	249
Weighted avg	0.25	0.50	0.34	249

TABLE VII: QNN Classification Report

C. Comparative Analysis of QML vs Classical ML

To effectively compare QML with classical ML, this work uses small sample sizes that comprise N rows and N features.

Category	**Model**	**Acc. (%)**
Classical	logistic regression	62
	SVM	71
	XGBoost	99
	random forest	92
Quantum	VQC	49
	QNN	50
	QSVM	47

Note: Evaluated on 846 samples with 29 features (846×29).

TABLE VIII: **Classical and Quantum Algorithm Accuracy Comparison**

Category	**Model**	**Acc. (%)**
Classical	logistic regression	66
	SVM	72
	XGBoost	89
	random forest	84
Quantum	VQC	48
	QNN	51
	QSVM	57

Note: Evaluated on 286 samples with 17 features (286×17).

TABLE IX: **Classical and Quantum Algorithm Accuracy Comparison**

Category	**Model**	**Acc. (%)**
Classical	logistic regression	64
	SVM	69
	XGBoost	67
	random forest	80
Quantum	VQC	52
	QNN	51
	QSVM	52

Note: Evaluated on 286 samples with 14 features (286×14).

TABLE X: **Classical and Quantum Algorithm Accuracy Comparison**

Category	**Model**	**Acc. (%)**
Classical	logistic regression	83
	SVM	83
	XGBoost	92
	random forest	83
Quantum	VQC	42
	QNN	50
	QSVM	58

Note: Evaluated on 51 samples with 14 features (51×14).

TABLE XI: **Classical and Quantum Algorithm Accuracy Comparison**

D. Interpretation and Optimization

In this section, we discuss and compare the performance of classical and QML algorithms in the NYPD crime dataset, highlighting key differences and performance outcomes. Classifiers, as supervised ML algorithms, can be essential in making quick decisions and allocating the appropriate personnel to efficiently address incidents, thus helping to predict the type of assistance required for each event or complaint call.

Classical ML algorithms currently represent the most practical and commercially viable approach to crime prediction. In our experiments, these classical models produced strong results: SVM achieved approximately 69-83% accuracy, logistic regression reached 62-83%, while XGBoost achieved

Category	Model	Acc. (%)
Classical	logistic regression	100
	SVM	100
	XGBoost	100
	random forest	100
Quantum	VQC	67
	QNN	33
	QSVM	67

Note: Evaluated on 51 samples with 7 features (51 × 7). Classical models overfit due to simple data structure; QSVM and VQC perform better, QNN poorly.

TABLE XII: **Classical and Quantum Algorithm Accuracy Comparison**

around 67-70%, and the random forest attained 82-90%. These results underscore the reliability and effectiveness of classical methods when working with large datasets and smaller and more constrained samples.

Optimal feature selection further enhanced the performance of classical models, particularly ensemble algorithms such as random forest and XGBoost. These methods inherently perform internal feature selection, allowing them to prioritize the most informative features and thereby improve predictive accuracy. Based on the results, the best performing models for classification tasks are the ensemble methods, namely, XGBoost and random forest Classifier. logistic regression proved to be too simple to capture the complex structure of the data, while SVM performed fairly well. Notably, SVM achieved up to 83% accuracy on a reduced dataset with fewer features, indicating its scalability. Regarding feature selection, the features that enhanced performance in classical algorithms had a similar, though less pronounced, effect on quantum models. However, this improvement was not statistically significant. This could be attributed to the current developmental stage of quantum algorithms, which are not yet fully optimized for handling complex feature interactions in practical applications. During the experiments, the VQC did not scale significantly, maintaining an accuracy range of approximately 48–50%. In contrast, the QSVM demonstrated some level of scalability, similar to its classical counterpart, achieving a maximum accuracy of 58%. The QNN provided a fair baseline, consistently hovering around the 50% accuracy mark.

Other challenges include limited qubit availability, long circuit execution times, and difficulties in scaling quantum models. These factors currently restrict the practical application of QML in crime prediction. Nevertheless, quantum models offer theoretical advantages, such as the ability to represent data in high-dimensional Hilbert spaces, which could enable them to capture more complex relationships as the technology matures.

V. Conclusion

Despite the current performance gap between classical and quantum algorithms, this study provides a useful benchmark comparison between the two. Exploring quantum models in simulation reveals key performance issues and highlights areas that need optimization. Understanding the strengths and weaknesses of each approach supports the ongoing discussion of the practical use of quantum computing in data science.

This research shows that, while quantum models still face major practical challenges, they offer theoretical benefits worth investigating. Continued work will be important as quantum hardware advances. Future studies could run models on actual quantum devices such as IBM Quantum instead of simulators. Further testing with alternative sampling methods (e.g., SMOTE), better feature selection, and improved preprocessing may also boost performance. Exploring other quantum machine learning algorithms could offer additional insight.

In conclusion, while quantum computing is not yet a viable replacement for classical machine learning in real-world crime prediction, its potential for future disruption is strong. As advancements in quantum hardware and algorithm design continue, quantum models may become valuable assets in data science over the next decade.

References

[1] J. W. M. Keith Spencer, "Crime," https://www.thecanadianencyclopedia.ca/en/article/crime, 04 2015.

[2] C. B. Canada, "The rise of ai: Fraud in the digital age," 03 2024. [Online]. Available: https://www.canada.ca/en/competition-bureau/news/2024/03/the-rise-of-ai-fraud-in-the-digital-age.html

[3] N. Shah, N. Bhagat, and M. Shah, "Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention," *Visual Comp. for Indus., Biomedicine, and Art*, vol. 4, no. 1, p. 9, 2021.

[4] M. Azure, "What is computer vision." [Online]. Available: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-computer-vision#object-classification

[5] J. Schneider and I. Smalley, "Quantum computing," https://www.ibm.com/think/topics/quantum-computing, 08 2025.

[6] S. Sharma, "Qeml (quantum enhanced machine learning): Using quantum computing to enhance ml classifiers and feature spaces," 2020. [Online]. Available: https://arxiv.org/abs/2002.10453

[7] M. Grossi, N. Ibrahim, V. Radescu, R. Loredo, K. Voigt, C. von Altrock, and A. Rudnik, "Mixed quantum–classical method for fraud detection with quantum feature selection," *IEEE Transactions on Quantum Engineering*, vol. 3, p. 1–12, 2022. [Online]. Available: http://dx.doi.org/10.1109/TQE.2022.3213474

[8] scikit-learn developers, "sklearn.metrics.confusion_matrix," https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html, 2023, accessed: [Insert access date here].

[9] M. Hdaib, S. Rajasegarar, and L. Pan, "Quantum deep learning-based anomaly detection for enhanced network security," *Quantum Machine Intelligence*, vol. 6, no. 1, p. 26, 2024.

[10] A. N. K. MD. Yasar Arafath, "Quantum computing based neural networks for anomaly classification in real-time surveillance videos," *Computer Systems Science and Engineering*, vol. 46, no. 2, pp. 2489–2508, 2023. [Online]. Available: http://www.techscience.com/csse/v46n2/51649

[11] Scikit-learn, "Glossary of common terms and api elements." [Online]. Available: https://scikit-learn.org/stable/glossary.html#term-classifier

[12] T. O. C. Tutor, "Continuous vs discrete data." [Online]. Available: https://www.youtube.com/watch?v=cz4nPSA9rlc

[13] Y. Du, X. Wang, N. Guo, Z. Yu, Y. Qian, K. Zhang, M.-H. Hsieh, P. Rebentrost, and D. Tao, "Quantum machine learning: A hands-on tutorial for machine learning practitioners and researchers," *arXiv preprint arXiv:2502.01146*, 2025.