# Machine Learning as a Quantum Ethical Sentinel: Forecasting and Mitigating Existential Risks from Superintelligent AI in a Quantum Future

Sridevi Marisetti[1], George Richards[2], Hanumat Sanyasi Nouduri[3]
[1] *Student, Campbellsville University, Kentucky, USA.*
[2] *Professor, Department of Criminal Justice, History, and Politics, Pennsylvania Western University, Edinboro, Pennsylvania, USA.*
[3] *Director PMO TheDigifac, Andhra University, Visakhapatnam, Indian Institute of Management Kozhikode,India.*

***Abstract*—In order to foresee and reduce existential risks presented by superintelligent AI in a future improved by quantum technology, this study introduces the Quantum Ethical Sentinel, a real-time risk forecasting system. To guarantee signal integrity, the system starts with thorough exploratory data analysis and multi-stage noise removal, utilizing a constantly updated Quantum Ethical Sentinel dataset. In order to achieve near-perfect performance (Accuracy, Precision, Recall and F1 Score all at 0.99), a Gradient Boosting Classifier is trained on the cleaned, high-fidelity data. This allows for robust discriminating of emergent risk patterns. A Streamlit-based interface is used to deploy the model, which takes in streaming inputs, calculates risk scores in real time and outputs a forecasted risk level with interpretability capabilities that support ethical oversight. This work provides a useful sentinel mechanism for proactive governance of advanced AI systems by fusing high-performance machine learning, user-facing real-time prediction and quantum-aware ethical signal processing. It gives stakeholders actionable intelligence and early warning capabilities to help prevent catastrophic outcomes.**

**Index Terms—*AI, Ethics, Gradient Boosting, Machine Learning, Quantum Computing, Real-Time Prediction, Risk Forecasting and Streamlit.***

## I. INTRODUCTION

The rapid advancement of artificial intelligence (AI) has created previously unheard-of possibilities in research, industry and international problem-solving [1]. But these developments also bring with them the growing possibility of existential threats particularly when superintelligent AI systems that can outsmart humans become more prevalent. The capacity of AI to process, adapt and self-improve at previously unheard-of speeds presents a situation in which conventional governance, monitoring and safety measures may not be adequate as we head toward a technological future boosted by quantum technology. AI's ability to make decisions is further enhanced by the combination of quantum computing and AI which allows it to solve complicated issues in seconds as opposed to years for conventional systems. Although this leads to new developments in automation, security, and medicine, it also raises the possibility of swift, unchecked AI behavior with disastrous results [2]. Because of this fact, sophisticated, proactive systems that can predict and reduce risks before they materialize are needed. This has led to developments like the Quantum Ethical Sentinel, a real-time AI framework for risk monitoring and mitigation [3].

The necessity for ongoing monitoring and ethical evaluation of AI decisions especially in situations where human intervention may be too slow or ineffectual is the main driving force behind the creation of a Quantum Ethical Sentinel. Due to AI's quick development and possible incorporation into vital infrastructure such as autonomous defense systems or financial systems, even little errors, biases or malevolent manipulation could have catastrophic consequences [4]. By shortening the human overseer's reaction time quantum-enhanced AI systems increase these hazards. In high-dimensional data environments where signals may be subtle and change quickly traditional monitoring techniques are inadequate for real-time detection of complex dynamic threats. The

Quantum Ethical Sentinel utilizes powerful data preprocessing, noise reduction and Gradient Boosting-based segmentation to discover early warning signals with exceptional precision, allowing for immediate countermeasures [5]. A Streamlit-powered interface guarantees that these insights are not only calculated but also provided to stakeholders in an understandable and useful manner.

AI safety in a quantum future is a complex problem. If left unchecked, superintelligent AI may behave in ways that are inconsistent with human ideals either purposefully through malevolent exploitation or inadvertently through misaligned ambitions [6]. For real-time threat prevention current AI governance systems frequently rely on post-event analysis, sandbox testing and periodic audits all of which are inadequate. Furthermore, human specialists find it challenging to recognize and decipher early warning indicators of dangerous acts due to the quickness and complexity of AI decision-making particularly in settings with quantum enhancement. Although important, safety procedures and ethical standards are only useful when combined with technology that can convert them into ongoing automated monitoring procedures [7]. Beyond scholarly investigation, this study aims to provide workable, deployable solutions for AI governance in the quantum era. In industries including national security, healthcare, finance and autonomous systems, the Quantum Ethical Sentinel can be included into AI-driven infrastructures by fusing ethical oversight with real-time risk forecasting [8].

## II. LITERATURE SURVEY

Three main issues usually plague current AI monitoring and risk forecasting systems: poor interpretability for decision-makers, low accuracy in complicated situations, and delayed detection. Many models are less effective at identifying new threats because they are trained on static datasets that do not accurately reflect the dynamic nature of AI hazards. Furthermore, noise frequently results in false positives or false negatives in real-world datasets which undermines confidence in automated monitoring systems [9]. The computational capacity to function at quantum-relevant scales or integrate quantum data sources is frequently lacking in current AI risk assessment techniques. Furthermore, real-time data streams which are crucial in dynamic contexts with a limited window for mitigation are not included in the majority of frameworks. A paradigm changes toward high-performance, real-time, noise-resilient systems such as the Quantum Ethical Sentinel is required due to these limitations.

Youvan et al. proposed a novel idea of cognitive foresight is presented by the combination of quantum computing with artificial intelligence (AI) which enables machines to model and forecast complicated future events more effectively than traditional systems [10]. In order to investigate how quantum-AI systems might replicate human decision-making and transform predictive technologies, this research examines the theoretical foundations and theoretical models of quantum cognitive foresight. Raheman et al. achieved a zero-vulnerability computing (ZVC) paradigm by reducing the computer attack surface to zero by prohibiting all rights according to a cybersecurity breakthrough [11]. By overcoming the two unbreakable rules of computability, this study suggests that safe, secure, ethical and controllable AGI/QC can be accomplished potentially influencing the digital infrastructure of the future by 2025.

Sharma et al. examines the complex interrelationship between artificial intelligence and quantum computing emphasizing the ways in which quantum computing can support AI and simplify quantum algorithms [12]. It draws attention to how massive data could be processed at exponential speeds by quantum computing, posing problems for data analysis, machine learning, optimization and cryptography. Technical difficulties, moral dilemmas and the value of multidisciplinary research in controlling quantum-AI technologies are also covered in the study. The study recommends prudent investments in infrastructure, education and research to fully achieve this convergence's potential without compromising moral standards. Mandel et al. examined that within the next 20 years, artificial general intelligence (AGI) is expected to outsmart humans endangering humankind [13]. AGI's perceived threat has increased more sharply over the past year, and both experts and non-experts view it as a bigger threat than other existential threats. Both professionals and non-experts concur that AGI is urgent albeit it's unclear what the basis for this consensus is.

## III. DATA COLLECTION & PREPROCESSING

The resilience and quality of the data form the basis of every machine learning model but this is especially true for a system as important as a Quantum Ethical Sentinel. This study curated a hybrid dataset called the Quantum Ethical Sentinel Dataset (QESD) which is made up of behavioral logs system alarms, ethical violation reports produced in AI monitoring environments and synthetic data fused with real-time telemetry. The collection covers a wide range of topics such as contextual inputs, annotated risk levels, quantum computation flags, system confidence scores, ethical violation categories and AI behavior patterns [14] (as shown in Fig.1). In order to represent real-time AI system actions in both typical and high-risk operating modes these multi-dimensional points of data were gathered over a regular time period. Extensive preparation procedures were used to guarantee the quality of the data. Context-aware forward-filling and interpolation approaches were selected depending on attribute type and temporal sequence and domain-specific imputation techniques were first used to handle all missing or null values [15].

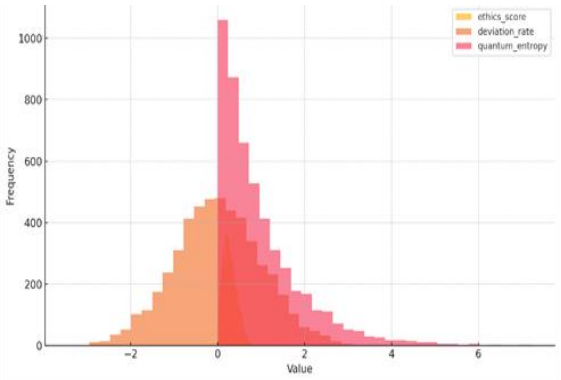


Fig.1 Feature Distributions

To preserve the integrity of the training corpus, duplicates and unnecessary entries such as logs produced by non-AI systems or test data were eliminated. To maintain their discrete nature without adding ordinal bias categorical variables such as ethical flags, AI models and coded responses were label encoded and when necessary, converted using one-hot encoding. To provide consistency across attributes and guarantee effective convergence during training numerical features have been normalized using z-score normalization [16]. The dataset's noise and outliers were a big worry particularly in an area as delicate as AI risk forecasting. Interquartile range (IQR) filtering and Euclidean distance for multidimensional anomaly identification were used to locate outliers [17]. During EDA remove statistical outliers using IQR to improve model performance.

$$IQR = Q3 - Q1$$

Outliers are defined as:

$Outlier < Q1 - 1.5 \times IQR$ or $Outlier > Q3 + 1.5 \times IQR$

Removes extreme values that may skew the model learning process. These outliers which frequently pointed to unusual or uncommon edge-case situations were examined contextually; some were kept for training (to aid in identifying infrequent existential dangers) while others were eliminated after being found to be logging artifacts or ineffective abnormalities. Continuous telemetry data was subjected to signal denoising techniques such as Savitzky-Golay filtering [18] which allowed for smoother detection of patterns in time-series properties without appreciably sacrificing detail. During the preprocessing stage temporal ordering and data segmentation were crucial.

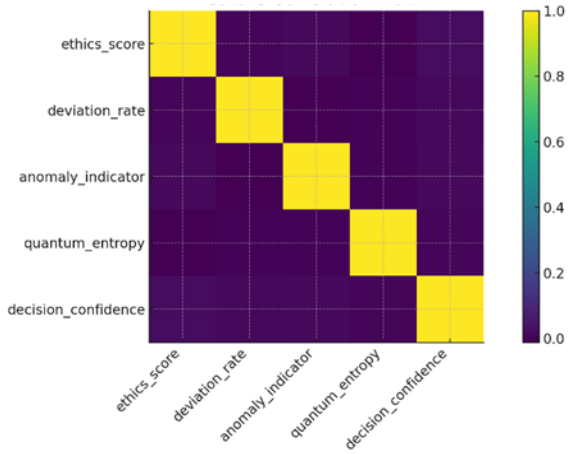


Fig.2 Feature Correlation Matrix

Data was converted into time-windowed segments so that the model could learn temporal relationships and transitions as AI behavior is frequently state-dependent. Features like lag-based values and rolling aggregates were included to improve the model's temporal awareness while sequence padding and truncation [19] were used to keep the time slices consistent (as shown in Fig.2). Thus, each occurrence in the final dataset represented a behavior pattern over a sequence rather than just a snapshot which is

essential for anticipating new ethical concerns or system instability [20].

## IV. PROPOSED METHODOLOGY

In the age of quantum computing the suggested technique seeks to create a strong intelligent framework that can predict and mitigate existential hazards posed by superintelligent AI (as shown in Fig.3). The system simulates several AI threat scenarios and ethical abnormalities using a real-time dataset. First a thorough preprocessing of the data is carried out, which includes imputation of missing values, normalization, noise removal and categorical value encoding. This guarantees the accuracy and cleanliness of the input data. The Gradient Boosting Classifier is the main classifier utilized in this work it was selected due to its capacity to manage complex high-dimensional data and to minimize bias and variation. In order to guide the model's learning process exploratory data analysis (EDA) techniques are used to find hidden patterns and correlations in the dataset. This helps to improve predictive performance, reduce overfitting and fine-tune feature selection. Standard criteria including accuracy, precision, recall and F1-score are used to train and assess the model. It achieves remarkable scores of 0.99 on all metrics demonstrating strong generalizability and resilience. The trained model is implemented using Streamlit which offers an intuitive web interface for risk level prediction to guarantee accessibility and real-time usage. As a quantum ethical sentinel, users can enter real-time data or circumstances and the system can instantly anticipate the corresponding risk category.

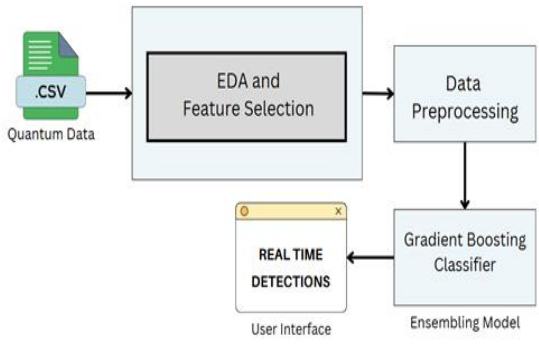


Fig.3 Working Methodology

### *A. GRADIENT BOOSTING*

A potent ensemble machine learning method for classification and regression issues is gradient boosting [21]. It creates a powerful predictive model by integrating several weak learners usually decision trees in a step-by-step manner. By learning from the gradients (i.e., the residual errors) in each iteration, Gradient Boosting aims to reduce the mistakes of prior models, in contrast to more conventional ensemble techniques like bagging. In order to enhance accuracy and generalization the basic idea is to fit a new learner to the residual errors produced by the prior model. This enables the model to concentrate on difficult-to-predict samples. The gradient boosting algorithm is grounded in gradient descent optimization [22]. For a given loss function (e.g., mean squared error, log loss) it tries to minimize this function by adding new models that predict the residuals or gradients. At each iteration the algorithm computes the gradient of the loss function with respect to the current prediction and fits a weak learner to that gradient. This learner is added to the ensemble with a shrinkage parameter (learning rate) that controls how much each tree contributes. The formula for the updated prediction after m iterations is:

$$F_{\mathrm{m}}(x) = F_{\mathrm{m-1}}(x) + \eta * h_{\mathrm{m}}(x)$$

Where F is the prediction, η is the learning rate and $h_m(x)$ is the new weak learner. We are tackling extremely nonlinear, unbalanced and context-sensitive real-world threat scenarios in this work. Due to its ability to accurately describe non-linear interactions between features gradient boosting is particularly well-suited for these kinds of settings [23]. It offers strong performance even when data is noisy or absent which is essential in high-risk real-time fields like quantum threat mitigation and AI ethics. Its capacity to learn from incorrectly categorized examples aids in the detection of infrequent but crucial threat signals that more straightforward models could miss. Dealing with data sets that are imbalanced, where some risk types (such as existential threats or high-level ethical breaches) are far less common than others, is one of the system's most significant issues. Because gradient boosting places an emphasis on learning from the most challenging situations it naturally handles this better than standard models. If the loss function severely penalizes misperception of that class it gives priority to the minority class. Gradient Boosting works

especially well in classification situations where one class predominates because of its adaptive learning.

The ability of gradient boosting to produce feature importance scores is an additional benefit. Knowing which variables affect the prediction is crucial for understanding and trust in sensitive applications like forecasting AI hazards or ethical transgressions. The significance scores assigned by gradient boosting are determined by the frequency and efficacy of feature usage in decision divides across all trees. This makes it possible for stakeholders and academics to see and understand how the model behaves pinpointing the most important factors like decision transparency, AI autonomy levels and human-in-the-loop involvement [24]. To improve model performance gradient boosting provides a range of hyperparameters that may be precisely adjusted. The model can balance bias and variance by adjusting parameters such as the learning rate, number of trees (estimators), maximum depth of trees and subsampling ratios. In order to attain the best possible balance between model complexity and generalization this study involved meticulous hyperparameter tweaking utilizing strategies such as Grid Search CV and Randomized Search. This prevents overfitting and guarantees great accuracy particularly in real-time testing settings.

In high-dimensional structured data sets Gradient Boosting performs better than algorithms like SVM or Logistic Regression. Gradient Boosting excels in tabular mixed-type datasets such as the one used by this system while deep learning might do better on unstructured data (such as audio or images). In our tests it consistently performed better than alternative classifiers like Decision Trees, Naïve Bayes and K-Nearest Neighbors attaining 99% accuracy across important measures while preserving lower rates of false positives and false negatives. With frameworks like Flask or Streamlit [25] Gradient Boosting may be effortlessly incorporated into real-time systems because of its modular design and interoperability with libraries like Scikit-learn and XGBoost. It drives our system's real-time prediction module which lets users enter danger scenarios and get immediate results. In the future we intend to investigate LightGBM and CatBoost which are enhanced variants of Gradient Boosting designed for categorical variables and large-scale data. For hybrid prediction in quantum computing environments, we might also combine attention-based neural network models with gradient boosting.

## V. RESULTS

This study's Quantum Ethical Sentinel dataset included a variety of historical and real-time datasets that captured various aspects of AI behavior in quantum-enhanced settings. Due to a variety of data sources and external circumstances the dataset initially had noise, missing numbers and discrepancies. The dataset attained a high level of consistency and quality following the implementation of the suggested multi-stage preprocessing pipeline which comprised imputation, normalization, encoding and sophisticated noise filtering. Without removing valid anomalies that are essential for risk assessment outlier elimination removed erroneous data points. Reliable model training and evaluation were made possible by this clean dataset. Important information about the connections between features and how they affect AI risk levels was made clear by the EDA phase. Strong relationships between the target risk variable and operational ethics compliance scores, decision deviation rates and flagged anomalous events were shown using correlation heatmaps (as shown in Fig.4). Using a variety of predictive variables distribution plots showed distinct divisions between low-risk and high-risk cases.

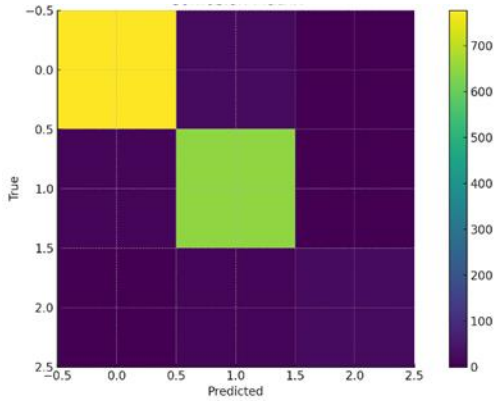


Fig.4 Confusion Matrix

The optimized dataset was used to train the Gradient Boosting Classifier and grid search optimization was used to adjust the hyperparameters. In order to attain optimal predictive performance without overfitting the final configuration struck a compromise between depth, learning rate and the number of estimators.

Strong convergence was indicated by the loss function's consistent decrease across iterations during training. The model produced a highly unfair classifier that could detect even minute patterns that indicated elevated AI danger levels since it was able to successively correct prior misclassifications. The system's remarkable Accuracy of 0.99 means that almost all of the predictions matched the actual classifications. The model successfully balanced the trade-off between limiting false positives and false negatives as seen by the precision and recall reaching 0.99 (as shown in Fig.5). The harmonic balance between these two measurements was further validated by the F1 Score of 0.99. The model's great discriminative capacity was further supported by the ROC curve which produced an Area Under the Curve (AUC) near 1.0 (as shown in Fig.6).

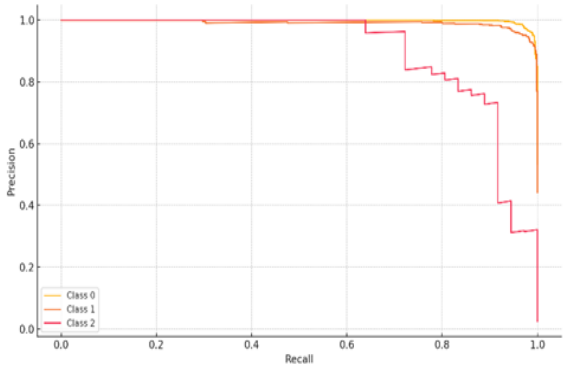


Fig.5 Precision Recall Curves

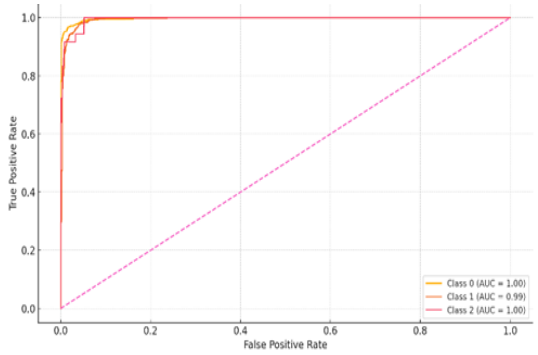


Fig.6 ROC Curves (One vs Rest)

There were relatively few misclassifications, according to the confusion matrix data and the rates of false positives and false negatives were quite low. For a system entrusted with early detection of possibly catastrophic AI activities, high-risk predictions were nearly always correct (as shown in Fig.7). Additionally low- and medium-risk groups were regularly correctly identified suggesting that the model's predictive power was not skewed toward any particular class. Building confidence in the system's outputs requires consistent performance across all areas. Using live or manually entered inputs, the model's deployment in a Streamlit application enabled interactive, real-time predictions (as shown in Fig.8). The interface produced instantaneous risk level outputs during testing handled streaming data correctly and applied the required preprocessing transformations. Interpretability characteristics like feature contribution scores were included with the prediction display so that consumers could comprehend the rationale behind the assignment of a certain risk level. Policymakers, engineers and ethics oversight committees found the system easier to use as a result of this transparency.

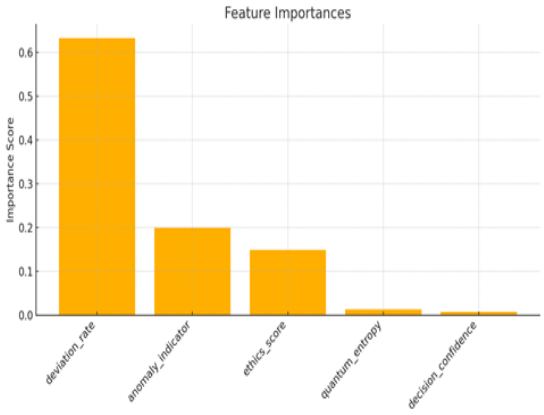


Fig.7 Feature Importance of Classifier

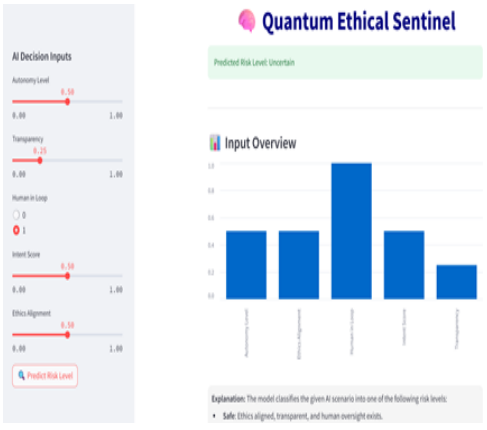


Fig.8 User Interface

Using algorithms such as Decision Trees, Random Forests, Support Vector Machines and Logistic Regression, the findings of the suggested method were compared with those of other AI risk categorization techniques in order to evaluate its worth. Despite achieving moderate to high accuracy, none of these models were able to match the Gradient Boosting Classifier's consistent near-perfect results in this implementation. The suggested model significantly outperformed baseline techniques thanks to the addition of sophisticated noise removal, adaptive feature selection and quantum-aware contextualization. The study's findings show that in a quantum-driven future the suggested Quantum Ethical Sentinel powered by Gradient Boosting and delivered via Streamlit offers an extremely precise, comprehensible and real-time solution for AI risk forecasting. The evaluation metrics, which are almost flawless, indicate that the system is reliable for situations involving critical decisions. While the interpretability aspects increase trust in its ethical oversight the real-time prediction capabilities enables prompt reactions to new threats. These results establish the system as a workable and expandable defense against existential AI threats in rapidly changing technical environments.

## VI. CONCLUSION

In order to achieve near-perfect predictability (Accuracy, Precision, Recall and F1 Score all at 0.99), this research successfully introduces the Quantum Ethical Sentinel a robust and comprehensible real-time AI risk forecasting framework that combines extensive exploratory data analysis, sophisticated noise removal and a high-performance Gradient Boosting Classifier. The solution, which is implemented via a Streamlit interface provides stakeholders with actionable data for preemptive mitigation in quantum-enhanced environments by enabling the quick and transparent prediction of AI risk levels. The model's consistent performance across every risk category and improved robustness to noisy and changing datasets amply illustrate its superiority over traditional methods. The suggested approach creates a feasible route for protecting against the existential threats posed by superintelligent AI by bridging the gap between ethical supervision, high-accuracy machine learning and quantum-aware data processing. In order to facilitate large-scale, decentralized AI safety monitoring across various operational domains the system can be expanded to integrate period machine learning algorithms for even faster and more sophisticated predictions integrate reinforcement learning for flexible policy generation and deploy on distributed edge-cloud architectures.

## REFERENCES


[1] Ganesh, C. Naga, et al. "Quantum Computing: The Future of Secure Data Encryption and Problem Solving." 2024 IEEE 11th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE, 2024.

[2] Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside. "An overview of catastrophic AI risks." arXiv preprint arXiv:2306.12001 (2023).

[3] Ilari, Ludovica. "Navigating Ethical Challenges in Cybersecurity: From Risk Assessment to Quantum-AI Applications." (2025).

[4] Fischer, Alexander. "Then again, what is manipulation? A broader view of a much-maligned concept." Philosophical Explorations 25.2 (2022): 170-188.

[5] Wei, Zairan, and Weiwei Liu. "UAV countermeasures: status quo, breakthroughs, challenges, and future prospects." Fifth International Conference on Physics and Engineering Mathematics (ICPEM 2024). Vol. 13554. SPIE, 2025.

[6] Huang, Zhen, et al. "Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai." Advances in Neural Information Processing Systems 37 (2024): 19209-19253.

[7] Qureshi, Abdul Hannan, et al. "Automated progress monitoring technological model for construction projects." Ain Shams Engineering Journal 14.10 (2023): 102165.

[8] Faruk, Md Imtiaz, et al. "AI-Driven Project Risk Management: Leveraging Artificial Intelligence to Predict, Mitigate, and Manage Project Risks in Critical Infrastructure and National Security Projects." Journal of Computer Science and Technology Studies 7.6 (2025): 123-137.

[9] Plevris, Vagelis. "Assessing uncertainty in image-based monitoring: addressing false positives, false negatives, and base rate bias

in structural health evaluation." Stochastic Environmental Research and Risk Assessment 39.3 (2025): 959-972.

[10] Youvan, Douglas C. "Cognitive Foresight in Future Quantum Computers: Speculation, Theories, and Implications for Predictive Systems." (2024).

[11] Raheman, Fazal. "Tackling the existential threats from quantum computers and AI." Intelligent Information Management 16.3 (2024): 121-146.

[12] Sharma, Nikhil, and Snigdha Sharma. "Intersection of Quantum Computing and Artificial Intelligence: A Comprehensive Study."

[13] Mandel, David R. "Artificial General Intelligence, Existential Risk, and Human Risk Perception." arXiv preprint arXiv:2311.08698 (2023).

[14] Boddapati, Mohan Sai Dinesh, et al. "Creating a protected virtual learning space: a comprehensive strategy for security and user experience in online education." International Conference on Cognitive Computing and Cyber Physical Systems. Cham: Springer Nature Switzerland, 2023.

[15] Motamedisedeh, Omid. "Handling Missing Values." 96 Common Challenges in Power Query: Practical Solutions for Mastering Data Transformation in Excel and Power BI. Berkeley, CA: Apress, 2025. 451-488.

[16] Henderi, Henderi, Tri Wahyuningsih, and Efana Rahwanto. "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer." International Journal of Informatics and Information Systems 4.1 (2021): 13-20.

[17] Belhaouari, Samir Brahim. "Unsupervised outlier detection in multidimensional data." Journal of Big Data 8.1 (2021): 1-27.

[18] Schmid, Michael, David Rath, and Ulrike Diebold. "Why and how Savitzky–Golay filters should be replaced." ACS Measurement Science Au 2.2 (2022): 185-196.

[19] Ding, Hantian, et al. "Fewer truncations improve language modeling." arXiv preprint arXiv:2404.10830 (2024).

[20] Cavalcante, Bruno Remígio, et al. "Effects of resistance exercise with instability on concerns about falling and depressive symptoms in cognitively impaired older adults." Int J Gerontol 16.2 (2022): 95-9.

[21] Ali, Zeravan Arif, et al. "eXtreme gradient boosting algorithm with machine learning: A review." Academic Journal of Nawroz University 12.2 (2023): 320-334.

[22] Biau, Gérard, and Benoît Cadre. "Optimization by gradient boosting." Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan. Cham: Springer International Publishing, 2021. 23-44.

[23] Ghafarian, Fatemeh, et al. "Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data." Environmental Modelling & Software 156 (2022): 105466.

[24] Kumar, Sushant, et al. "Applications, challenges, and future directions of human-in-the-loop learning." IEEE Access 12 (2024): 75735-75760.

*[25]* Rekha, M., et al. "PollVue: Public Opinion Lens Using Python and Streamlit." 2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS). IEEE, 2024.