Quantum Computing: A Tool in Big Data Analytics

AKSHAT GAURAV[1], KWOK TAI CHUI[2], FRANCESCO COLACE[3],
[1]Ronin Institute, Montclair, USA, Email: akshat.gaurav@ieee.org
[2]Hong Kong Metropolitan University (HKMU), Hong Kong, Email: jktchui@hkmu.edu.hk
[3]University of Salerno, Italy, Email: fcolace@unisa.it

ABSTRACT Massive volumes of data, referred to as "big data," can do amazing things. Because of the enormous potential it has, it's been a hot issue for the previous two decades. Public and commercial sector organisations are using big data analytics to better the services they provide. Data management and analysis are necessary in order to extract relevant information from this little amount of data. Instead of searching for answers in vast data, it becomes like searching for the needle in the haystack. As a result of its many promises for information processing systems, quantum computing comes to the rescue, notably in the area of Big Data Analytics. Quantum computing's power is built on quantum physics' principles. Because these events lack a classical counterpart, conventional computers is unable to provide the same results. Here, we've taken a look at what's out there on Big Data Analytics using Quantum Computing. As a completely new subject, quantum computing presents a number of open issues. Quantum computing in Big data analytics is also emphasised for its problems, potential, and future directions and methodologies.

KEYWORDS Quantum computing, Big data

I. INTRODUCTION

Unlike traditional computing, quantum computing has its origins in a variety of areas. In the subject of information processing, it's an application of quantum physics ideas. In comparison to conventional computation, quantum computing provides fundamentally new answers to computational issues and makes problem-solving more efficient. Using this approach in 1994 sparked the "big bang" of quantum calculations, paving the path for the creation of quantum computing and the assessment of quantum computers [1]–[3]. Three quantum resources, none of which have mirror representations in conventional processing, provide quantum computing its potential capability. A function may be computed on an infinite number of inputs concurrently using the Quantum Parallelism concept of superposition with linearity in quantum mechanics. Quantum interference enables logical channels of a computation to interfere constructively or destructively, leading to desirable outcomes by strengthening one another and unwanted ones by cancelling one another. It's impossible to explain the function of multi-particle quantum states using a single state for each particle. Figure 1 depicts a hypothetical diagram of quantum computing's development [4]. There is a functional overlap between classical and quantum computing, but the physical layer is fundamentally distinct. Quantum computing technologies are based on the DiVincenzo criteria, which are complemented with specific physical layer features. Qubits, quantum registers, gates, circuits and memories are all produced from quantum computing technology under certain circumstances.

In the modern world, information has been the driving force for greater organisation and innovative ideas. We can better arrange ourselves to get the greatest results when we have more knowledge. As a result, gathering and analysing data is critical for any business. This data may also be used to make predictions about the present trends in various parameters and about what will happen in the future. Our growing awareness of this has led us to begin gathering and collecting more data on almost everything via the introduction of technological advancements. As a society, we're inundated with information on every element of our lives, from social interactions to scientific discoveries to professional careers to personal health. There are some parallels to be seen between the current situation and the recent data deluge. As technology has advanced, we've been able to generate ever more data, to the point that it's now impossible to handle with the tools we have. As a result, the phrase 'big data' was coined to characterise data that is enormous and unwieldy. Large-scale data analysis is a hot topic in academia right now and will continue to be in the future. As part of its annual "Top 10 Strategic Technology Trends For 2013" and "Top 10 Critical Technology Trends For The Next Five Years" lists, Gartner included Big Data in each of these categories [5], [6]. Yes, it is correct to suggest that Big Data will have a profound impact on numerous industries, from business to scientific research to governmental administration. Ten Vs may be used to describe the complexity of big data: Volume,

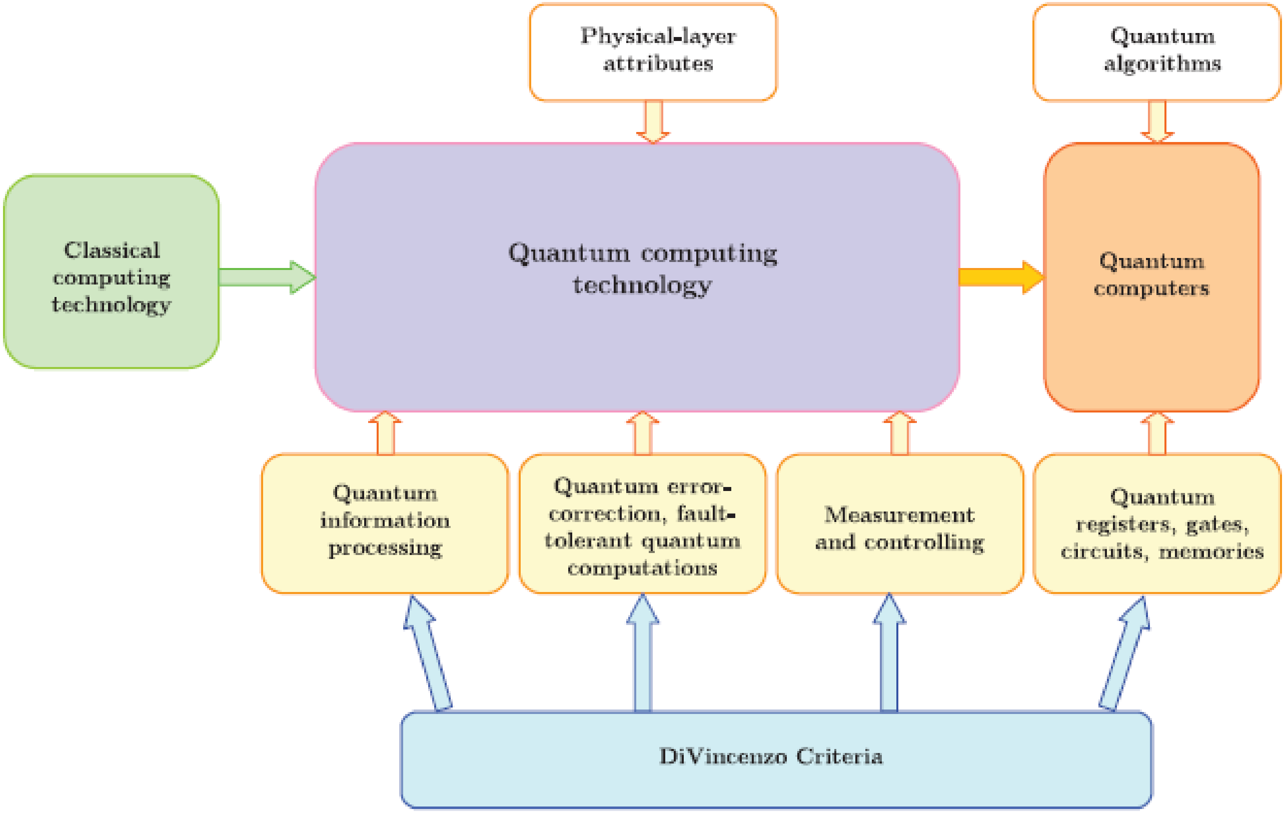
Physical-layer attributes
Quantum algorithms
Classical computing technology
Quantum computing technology
Quantum computers
Quantum information processing
Quantum error-correction, fault-tolerant quantum computations
Measurement and controlling
Quantum registers, gates, circuits, memories
DiVincenzo Criteria

FIGURE 1: Evaluation of Quantum computing [1]

Variety, Validity, Viability, Volatility and Vulnerability can all be represented by the 10 Vs. The following is an explanation of each 'V' in health care big data.

- In big data analytics technology, value indicates the significance of the data analysed. Processing medical data using an inaccurate processing approach diminishes the value of the processed data.
- The created data's volume is represented by this value. Health care information is complicated, and at times it includes a significant amount of noise. Patient records, biometric sensor readings, and x-ray pictures are among the most common types of healthcare data. Global healthcare data was estimated to be 500 petabytes in 2012 and is predicted to become 2500 petabytes by the end of 2020, according to the estimate. 1.2 to 2.4 Exabytes of healthcare big data are created each year.
- Velocity - This is the pace at which healthcare data is processed, i.e., the rate at which data is generated, stored, and then transported. There are several sites where healthcare data is created, including laboratories and doctor's offices. There must be a quick data interchange between multiple sites for real-time analysis of large healthcare datasets
- Veracity is a measure of the data's dependability. There is a greater chance that a patient's life will be saved if the data is accurate. This data should be gathered from many sources and processed using error-free technologies in order to boost its authenticity.
- Structured, semi-structured, and unstructured data are all types of healthcare data that are obtained from a variety of sources. Hence, diversity depicts the wide range of variations that exist in healthcare information. Healthcare data processing provides a significant problem.
- It refers to the accuracy of the medical data. When it comes to healthcare data, validity is critical since if the data isn't legitimate, the processed results are useless. There are a variety of methods and instruments for verifying the accuracy of healthcare data.
- When it comes to locating the essential information in a big ocean of data, the process becomes much more difficult since the amount of healthcare information is so enormous. In order to utilise healthcare data effectively, we must first pick relevant data.
- An enormous quantity of healthcare data is created every second because to the digital revolution. In addition to the previously saved data, this new data is also included. The next step is to determine the data's life expectancy. Healthcare data loses relevance with time, making it a research topic to determine the appropriate lifespan of the data.
- An important goal in healthcare is to create an efficient

system that can handle a significant volume of healthcare data in the most efficient manner possible. As a result, less effort is put into finding ways to make the healthcare system more susceptible to cyberattacks. The healthcare system begins to lose its significance as soon as it becomes exposed to security assaults.

- When it comes to big data, one of the most important things to keep in mind is that it has to protect the privacy of patients while still providing the information they need. Visualization is an effective tool for accomplishing both of these objectives.

The pace of data production accelerated with the dawn of the new millennium. One-fifth of Google's 1.2 trillion searches per year are first-time queries. From 2 zettabytes, the value of big data has more than doubled in the last decade, reaching 59 zettabytes. Big data has grown by 90% in the previous two years, according to an IBM research in 2017 [7]–[10]. Every day, internet users create over 2.5 trillion bytes of data. The total quantity of data created in 2020 will be dominated by North America, although diverse nations will also make significant contributions. Industry and service industries of all sizes contribute to the growth of big data. Major corporations like IBM, Oracle, Amazon, and Google, among others, are spearheading big data initiatives in order to deal with the influx of information. The fact that governments from various nations are interested in the development of big data projects shows how important big data is to the advancement of the technology [11]–[14]. Big data analytics is attracting the interest of the healthcare industry because of its ability to handle large amounts of data. The healthcare industry creates a lot of data, and by the end of 2020, it is predicted that the entire quantity of data created by the sector would be over 35ZB. Due to this importance of Big data, we focus on the reviews for big data analytics techniques [2], [15]. It's possible that the citation networks we've found in the scientific literature might show emerging trends and cooperation networks in the field of quantum computing and big data analytics. Researchers may look for the most well-known academic articles, grant applications, data sets, and patents in the Lens database [16]. As a consequence, we acquire our article entries and the associated citation network from the Lens database. "TS1=("Quantum Computing")" and "TS2= (("Quantum Computing") AND "Big data")" are used to search the Lens database [17] for relevant publications in quantum computing and big data analytics. The data spans the period from 2010 to 2021. When we finished setting up the various quarries, we received the various pieces of paper seen in Figure 2. Quantum computing and big data analytics based on quantum computing are flourishing topics, as indicated by the exponential rise in publications over the preceding decade, as seen in Figure 2. It's time for a comprehensive review in this field to help researchers keep up with the latest developments in the research problem.

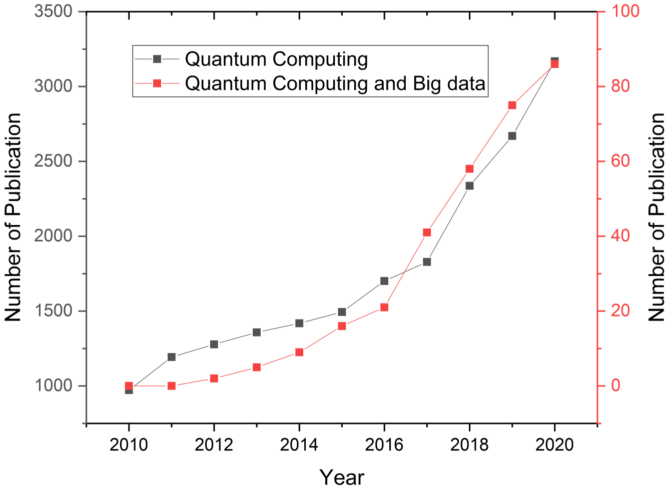

FIGURE 2: Number of Document Published for Quantum Computing

II. QUANTUM COMPUTING STATE OF ART

The rules of quantum mechanics are used to carry out calculations in quantum computing. It is at the molecular level and below that quantum mechanics comes into play, since it is the governing physical theory of all matter. A wave-like property is a property of both particles and waves, according to this theory. In contrast to a conventional computer that uses randomness and may produce a wide range of results, a quantum machine can have a wide range of amplitudes along its computational routes, much as a wave could. This wave-like activity might be harnessed for computing gain in the same way that random choices could lead to diverse outcomes in a classical computer. As soon as you start measuring, the state will 'collapse,' and you'll get an answer with a probability equal to amplitude sqrt Because they enable interference between computational routes akin to the interference between waves6, quantum computers hold the promise of a new kind of computing that is fundamentally distinct from any prior (classical) computing form. Some tasks can be solved faster using a quantum computer than a conventional computer, despite the fact that the greatest general-purpose quantum computers currently exist contain just 50–100 qubits. Two-level systems, such as a photon travelling down one of two optical fibres, are referred information as "qubits." There are two ways to think about qubits: as a generalisation of classical bits (cbits), and as a way to think about the state of a single qubit. |a0|2 + ||a1|2 = 1. Quantum computers' strength stems from their ability to scale [16]. The state of n qubits is defined by a complex unit vector of dimension 2n, while the state of a n cbit system may be in any one of 2n potential states at any one moment. Many of these vectors (also called wavevectors or wavefunctions) may be easily converted by multiplying them with unitary matrices. With O(n2) basic quantum gates, a wavevector may be Fourier converted. As a result, certain changes can't be done effectively. This information can only be derived from a quantum state by following the rules of quantum measurement. With a chance of |ax|2 of the state being destroyed, a complete state measurement produces an outcome of x. As a result, even though the amount of information required to describe the quantum state of n qubits grows exponentially as n increases, measurements can only retrieve n bits of information. Finding a means to take use of quantum computers' exponential state space despite these and other limitations is the primary issue of quantum algorithm creation.

III. BIG DATA ANALYTICS CHALLENGES

The term "big data" refers to the massive volumes of data that are being created at a fast pace. It is more important to optimise customer services than it is to maximise consumer consumption of the data obtained from diverse sources. Big data from biomedical research and healthcare also falls into this category. Big data is a tremendous difficulty because of the sheer number of information it contains. As part of the scientific community, data must be kept in a manner that is readily accessible and usable for efficient analysis. Implementation of high-end computing tools, protocols, and high-end hardware in the clinical environment is another key difficulty in the context of healthcare data. We need experts from several fields to accomplish this aim, including biology and information technology as well as statisticians and mathematicians. Pre-installed software solutions provided by analytic tool makers may make sensor data accessible on a storage cloud. AI specialists have built data mining and machine learning features for these instruments, which would be able to turn data into knowledge. Implementation would improve the efficiency of healthcare big data collection and analysis as well as the display of such data. The primary goal is to annotate, integrate, and display this complicated material in a way that facilitates comprehension. Research in biomedicine is hampered by a lack of clear information about the data available. Finally, computer graphics designers have devised visualisation tools that can effectively convey this new information. Big data analysis must also contend with the problem of data heterogeneity. Big data in healthcare is difficult to make sense of because of its enormous quantity and very varied nature. High-power computer clusters accessible through grid computing infrastructures are the most prevalent platforms for executing the software framework that supports big data analysis. Because of its virtualized storage and dependable services, cloud computing has become a popular choice for businesses. Highly scalable, highly reliable, and completely self-sufficient are just a few of the benefits that come with this system. Such platforms may operate as a receiver of data from the omnipresent sensors, as a computer to analyse and interpret the data, and as a web-based visualisation tool for the user. Mobile edge computing cloudlets and fog computing may be used in IoT to process massive data closer to the data source. ML and AI methods to large data analysis on computer clusters need the use of advanced algorithms. Could be written in a programming language suited to coping with large amounts of data. As a result, handling the massive amounts of data generated by biomedical research requires both biology and IT expertise. Bioinformaticians often have a dual skill set like this.

IV. BIG DATA AND QUANTUM COMPUTING

An efficient linear and non-linear binary classifier may be implemented on a quantum computer using a support vector machine with exponential speedups in the size of the vectors and number of training instances [18]. Performing a principal component analysis and a matrix inversion effectively is at the heart of the approach, which relies on a non-sparse matrix simulation technique. Weinstein [19] discusses the strong visual approach of dynamic quantum clustering (DQC), which works with large and high-dimensional data. Because it uses differences in data density (in feature space) and uncovers hidden subsets, it can handle large, high-dimensional datasets. To demonstrate how and why data points are identified as simple cluster members with correlations among all variables assessed, a DQC analysis has produced a video. According to Rebentrost et al. [20], a quantum computer-implemented optimal binary classifier with logarithmic complexity in vector size and training example number has been developed. The Generalized Eigen value Proximal SVM (GEPSVM) was introduced by Marghny et al. [21] to address the SVM complexity problem. Real-world data is affected by errors or noise, and dealing with this data may be a difficult challenge. This issue has been addressed in this paper with the help of a solution. DSAGEPSVM is the name of this approach. Using Quantum Computing, Anguita et al [22] explored how to overcome the challenge of effective SVM training, particularly in the case of digital implementation. Experiments in synthetic and real-world scenarios are conducted to support the theoretical understanding of the behavioural characteristics of standard and improved SVMs. The study provided here examines the similarities and contrasts between quantum-based optimization and quadratic programming. Future research challenges in Big data and quantum comouting Companies are always working to create new methods for managing and analysing massive amounts of data in order to better incorporate that data into their operations. In spite of this, the wide variety of products available has made it difficult to share data. These are a few of the issues we'll touch on briefly in this section. Data storage is one of the main issues, although many firms are satisfied with storing their own data on their own premises. Control over security, access, and uptime are just a few of the benefits. Scaling and maintaining an on-premises server network may be costly and time consuming. Cloud-based storage employing IT infrastructure looks to be a more cost-effective and reliable alternative for most healthcare firms, according to this study. Organizations should only work with cloud service providers that are cognizant of the need of security. For these reasons and more, cloud storage is becoming increasingly popular. A hybrid approach to data storage may be the most adaptable and viable option for providers with different data access and storage demands. Cleaning

After collection, data must be cleaned or scrubbed to guarantee its precision, correctness, consistency, relevance, and purity. Automated logic rules may be used to achieve high levels of correctness and integrity in this cleaning procedure. Machine learning methods may be used to decrease costs and time, and to prevent bad data from derailing big data initiatives, using more complex and accurate tools. Managing large amounts of data is very challenging, particularly when imperfect data is involved. Data storage and processing and sharing necessitates the creation of a unifirma format. Protection breaches, hacks, phishing assaults and even ransomware attacks have made data security a top responsibility for every firm. A set of technological protections was built for the protected stored data after a number of vulnerabilities were discovered. Organizations are guided by these regulations, known as HIPAA Security Rules, when it comes to storage, transmissions, authentication procedures and controls over access, integrity, and auditing. Using the latest antivirus software, firewalls, encrypting sensitive data, and multi-factor authentication may save you a lot of time and money in the long term. Having comprehensive, accurate, and up-to-date metadata on all of the data stored is essential to a successful data governance strategy. The metadata would include information such as the date of creation, the purpose of the data, and who was accountable for it. Later scientific research and precise benchmarking might benefit from analysts being able to reproduce prior questions. As a result, data is more usable and "data dumpsters" are less likely to be created with useless data.

With the use of metadata, businesses would be able to query their data and come up with some answers. However, query tools may not be able to access a complete repository of data if datasets are not properly interoperable. Moreover, a full picture of a patient's health may not be created if distinct dataset components are not adequately integrated or linked and readily available. Visualizing data using charts, heatmaps, and histograms to show contrasts and precise labelling may make it much simpler for humans to absorb the information and apply it correctly. There are a variety of other examples of data visualisations, such as bar charts, pie charts, and scatterplots.

V. CONCLUSION

Scientific revolutions are set to enter a new phase because of Big Data's role in innovation, competitiveness and productivity. It is a good thing that we will be able to observe the technology leapfrogging in the near future. In this article, we provide a quick introduction to the fundamentals and challenges of Big Data and quantum computing. These technologies are still in the early stages of development but we are certain that we will see a number of major breakthroughs in the near future. Although Big Data analytics is still in the early stages of development, the current Big Data approaches and tools are unable to tackle all of the genuine Big Data challenges. Consequently, governments and businesses should devote more resources to this scientific paradigm in order to reap the benefits of Big Data. More study in these sub-fields is needed to handle the problem of Big Data ith the help of quantum computing.

REFERENCES

[1] B. B. Gupta, A. Gaurav, and D. Peraković, "A big data and deep learning based approach for ddos detection in cloud computing environment," in 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE). IEEE, 2021, pp. 287–290.

[2] C. L. Stergiou and et al., "Secure machine learning scenario from big data in cloud computing via internet of things network," in Handbook of computer networks and cyber security. Springer, 2020, pp. 525–554.

[3] B. B. Gupta, S. Yamaguchi, and D. P. Agrawal, "Advances in security and privacy of multimedia big data in mobile and cloud computing," Multimedia Tools and Applications, vol. 77, no. 7, pp. 9203–9208, 2018.

[4] L. Gyongyosi and S. Imre, "A survey on quantum computing technology," Computer Science Review, vol. 31, pp. 51–71, 2019.

[5] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," Information sciences, vol. 275, pp. 314–347, 2014.

[6] K. Yadav and et al., "2021 hot topics in machine learning research."

[7] F. J. G. Peñalvo, T. Maan, S. K. Singh, S. Kumar, V. Arya, K. T. Chui, and G. P. Singh, "Sustainable stock market prediction framework using machine learning models," International Journal of Software Science and Computational Intelligence (IJSSCI), vol. 14, no. 1, pp. 1–15, 2022.

[8] K. T. Chui and et al., "Enhancing electrocardiogram classification with multiple datasets and distant transfer learning," Bioengineering, vol. 9, no. 11, p. 683, 2022.

[9] D. Singh, "Captcha improvement: Security from ddos attack," 2021.

[10] A. Gaurav, V. Arya, and D. Santaniello, "Analysis of machine learning based ddos attack detection techniques in software defined network," Cyber Security Insights Magazine (CSIM), vol. 1, no. 1, pp. 1–6, 2022.

[11] B. Joshi and et al., "A comparative study of privacy-preserving homomorphic encryption techniques in cloud computing," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–11, 2022.

[12] R. K. S. Rajput, D. Goyal, A. Pant, G. Sharma, V. Arya, and M. K. Rafsanjani, "Cloud data centre energy utilization estimation: Simulation and modelling with idr," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–16, 2022.

[13] F. J. G. Peñalvo, A. Sharma, A. Chhabra, S. K. Singh, S. Kumar, V. Arya, and A. Gaurav, "Mobile cloud computing and sustainable development: Opportunities, challenges, and future directions," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–20, 2022.

[14] K. Pathoee and et al., "A cloud-based predictive model for the detection of breast cancer," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–12, 2022.

[15] B. B. Gupta, Modern Principles, Practices, and Algorithms for Cloud Security. IGI Global, 2019.

[16] P. S. Emani, J. Warrell, A. Anticevic, S. Bekiranov, M. Gandal, M. J. McConnell, G. Sapiro, A. Aspuru-Guzik, J. T. Baker, M. Bastiani et al., "Quantum computing at the frontiers of biological sciences," Nature Methods, vol. 18, no. 7, pp. 701–709, 2021.

[17] "The lens - free & open patent and scholarly search," https://www.lens.org/, accessed: 2023-01-01.

[18] P. Rebentrost, M. Mohseni, and S. Lloyd, "Quantum support vector machine for big data classification," Physical review letters, vol. 113, no. 13, p. 130503, 2014.

[19] M. Marghny, R. M. A. ElAziz, and A. I. Taloba, "Differential search algorithm-based parametric optimization of fuzzy generalized eigenvalue proximal support vector machine," arXiv preprint arXiv:1501.00728, 2015.

[20] D. Anguita, S. Ridella, F. Rivieccio, and R. Zunino, "Quantum optimization for training support vector machines," Neural Networks, vol. 16, no. 5-6, pp. 763–770, 2003.

[21] T. A. Shaikh and R. Ali, "Quantum computing in big data analytics: A survey," in 2016 IEEE international conference on computer and information technology (CIT). IEEE, 2016, pp. 112–115.

[22] R. Dridi and H. Alghassi, "Homology computation of large point clouds using quantum annealing," arXiv preprint arXiv:1512.09328, 2015.