Indonesian Journal of Electrical Engineering and Computer Science
Vol. 35, No. 1, July 2024, pp. 156~164
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v35.i1.pp156-164

Cloud-based machine learning algorithms for anomalies detection

Raveendra N Amarnath[1], Gurumoorthi Gurulakshmanan[2]
[1]Project Lead, Mphasis, Campbell Dr, Melissa, Texas, USA
[2]Project Manager, Mphasis, Texas, USA

Article Info

Article history:

Received Oct 31, 2023
Revised Mar 7, 2024
Accepted Mar 20, 2024

Keywords:

Bayesian networks
Gated recurrent units
Gradient boosting machines
Intrusion detection system
Word2Vec

ABSTRACT

Gradient boosting machines harnesses the inherent capabilities of decision trees and meticulously corrects their errors in a sequential fashion, culminating in remarkably precise predictions. Word2Vec, a prominent word embedding technique, occupies a pivotal role in natural language processing (NLP) tasks. Its proficiency lies in capturing intricate semantic relationships among words, thereby facilitating applications such as sentiment analysis, document classification, and machine translation to discern subtle nuances present in textual data. Bayesian networks introduce probabilistic modeling capabilities, predominantly in contexts marked by uncertainty. Their versatile applications encompass risk assessment, fault diagnosis, and recommendation systems. Gated recurrent units (GRU), a variant of recurrent neural networks, emerges as a formidable asset in modeling sequential data. Both training and testing are crucial to the success of an intrusion detection system (IDS). During the training phase, several models are created, each of which can recognize typical from anomalous patterns within a given dataset. To acquire passwords and credit card details, "phishing" usually entails impersonating a trusted company. Predictions of student performance on academic tasks are improved by hyper parameter optimization of the gradient boosting regression tree using the grid search approach.

Corresponding Author:

Raveendra N Amarnath
Project Lead, Mphasis, Campbell Dr, Melissa
Texas 75454, USA
Email: raveendra.techie@gmail.com

1. INTRODUCTION

The first step to addressing a system security problem is recognizing it. Intrusion detection systems scan network traffic for signs of assaults and are used to detect them. Intrusion detection finds these. An intrusion detection system (IDS) monitors network traffic for intrusions [1]. The majority of phishing assaults work because people have already developed trusting relationships with businesses [2]. Phishing involves impersonating a reputable firm to obtain valuable data including user names, passwords, and credit card numbers through email. Phishing attacks utilize more than fake emails and websites. Blacklists of hazardous uniform resource locators (URLs) and internet protocol (IP) addresses may be used to detect phishing websites [3]. Companies may keep their income steady by focusing on client retention rather than customer acquisition, which can be five to ten times more costly. The telecoms sector relies heavily on churn rate modeling and analysis [4]. Inadequate accuracy in identifying certain kinds of ambiguity is a consequence of the fact that present suggested strategies do not adequately treat all forms of linguistic ambiguity [5].

This ensemble model, which makes use of data from learning systems, supposedly provides more accurate results than conventional models and gives managers new ways to improve instruction and student outcomes [6]. Therefore, customer relationship management (CRM) must adapt to these developments in order to provide a competitive advantage to the company. The cost of a no satisfaction consumer is high since negative reviews and comments may quickly spread over the Internet, potentially reaching millions of people [7]. Profitable marketing strategies thatmaximize shareholder value focus on retaining existing customers and minimizing customer turnover. In the telecoms business, customers are still the most important asset. A drop in clientele is a shock to any company's finances [8]. It's the internet, but for the whole world. Host computers and server computers contain data that may be accessed over the internet. internet protocol/transmission control protocol (IP-TCP) was the protocol used for communication [9]. Cognitive decline makes it difficult to recognize the signs of depression in the elderly. A common tool for identifying depressive symptoms in the elderly is the geriatric depression scale (GDS) [10].

Online travel agencies (OTAs) are prominent e-commerce platforms that provide hotel bookings. Customers may help each other pick a hotel by submitting reviews [11]. English has received more attention in named entity recognition (NER) studies than any other language in part because of the abundance of high-quality annotation datasets in the language. Natural language processing (NLP) performance may be enhanced by the recognition of corpus entities, for which NER is crucial [12]. Classification can immediately analyze text in natural language processing. Text categorization may be used to assess question difficulty by making informed assumptions about unknown variable values based on other factors [13]. Instagram, like many other social media sites, has realized the critical need of hate speech identification. The usefulness of these techniques varies with the language and dataset used [14]. A large number of people from all over the globe visited Thailand before the spread of corona virus (COVID-19). The hospitality business must be ready, adapt, and present new services for guests after the COVID-19 pandemic. One of a company's most valuable resources nowadays is its data [15].

Recurrent neural networks (RNN) models outperform logistic regression, random forest, and support vector machines in comparable conditions. As a result, it seems that more sophisticated computational techniques are required for extraction of the complex underlying linkages [16]. As noted, as research in explains, the ontology creation or construction process is rather expensive since most of it is done manually. In light of these considerations, various researches have been carried out to automate the ontology building or creation process [17]. Yet, despite its usefulness, social media also presents challenges, such as the ease with which false information may be disseminated. Since the context analysis in a convolutional architecture is based on patterns across several texts, it excels in feature extraction but lacks the capacity to recall earlier knowledge. While recurrent architecture is designed to take information order into account, its feature extraction capabilities lag below those of convolutional architecture [18]. The accessibility of social media has contributed much to its meteoric rise in popularity in recent years. Many young people nowadays rely heavily on social media technologies because of the social connection and simple content production they provide [19]. Well-defined software needs specification (SRS) outlining the system's functional and non-functional requirements is essential to guarantee software systems deliver as promised. Investigating the edge/cloud application's issue SRS early ensures the final result satisfies audience expectations [20].

A unique thing, location, or person is a name entity (NE). Humans can recognize many self-named things with capital letters, but robots struggle with NER [21]. Internet of things (IoT) sensors, data analytics and artificial intelligence (AI), precision agriculture, remote monitoring and control, automated systems, livestock management systems, cloud computing and big data storage, energy management, and farm management software are now in smart farms [22]. Flood frequency has increased over the previous 30 years, devastating infrastructure. They supply critical goods and services and boost social and economic prosperity [23]. Natural disasters are expected to grow in frequency and severity as a result of climate change. Heavy rainfall, melting glaciers, and rising sea levels are all made more likely by extreme weather [24]. The different products offered by banks may be evaluated alone or in combination. The model will be built using empirical data that can be put to the test [25].

2. PROPOSED SYSTEM

There is a certain order of actions that must be followed while administering medical care to a patient. First, a complete evaluation of the patient is made, including a review of the patient's medical history, a physical examination, and any other necessary diagnostic procedures. After the evaluation is over, doctors will look at the results and make a diagnosis based on what they find. The next step is to devise a treatment strategy, which may include the selection of medication, therapy, or surgical procedures. During the therapy phase, patients are closely monitored, and the treatment plan is modified as needed to account for any problems or unwanted side effects. Maintaining open lines of communication and educating patients on their conditions, treatment plans, and the significance of sticking to their regimens is essential at every stage.

Multidisciplinary teams may need to work together to provide care, and effective communication is crucial for this to happen [26]. Treatment plans may be fine-tuned based on regular progress checks and reevaluation of the patient's condition. The ultimate objective of medical treatment is to improve health outcomes for patients.

The goal of research into cloud-based machine learning techniques and their many potential uses is to make the most of this disruptive technology in solving difficult issues in a wide range of fields. These procedures make it easier to implement complex algorithms on scalable cloud architectures, which speeds up the processing and analysis of massive data sets. Predictive analytics for illness diagnosis and treatment planning are made possible by machine learning algorithms hosted on the cloud, which improves the quality of care provided to patients. More than that, they help firms thrive by enhancing performance through demand prediction and consumer understanding. These techniques aid in environmental monitoring and resource management by facilitating data-driven decision making, which is essential for combating climate change and improving sustainability. Cloud-based machine learning enables algorithmic trading models that improve financial investing tactics. By facilitating real-time data processing for secure navigation, these techniques also support the growth of autonomous cars. These algorithms are also responsible for the success of other online services, such as tailored content suggestions and fraud detection in online platforms. As such, the goal is to realize the full potential of cloud-based machine learning algorithms across a wide variety of disciplines, therefore expanding our understanding and inspiring new approaches to solving complex problems.

Exploring cloud-based machine learning techniques and their applications is innovative because it focuses on using state-of-the-art technology to solve complex problems in many fields. The goal is to process and analyze huge, complicated information quickly and easily by taking use of cutting-edge algorithms hosted on elastic cloud computing platforms. Predictive analytics for better healthcare results, streamlined company operations, heightened environmental monitoring and resource management, and smarter financial strategies are just some of the ways this novel method hopes to alter established norms. The ultimate objective is to accelerate development in a wide range of disciplines by using the disruptive power of cloud-based machine learning to address intractable issues and fuel breakthroughs in previously unrelated areas.

2.1. Gradient boosting machines in cloud-based machine learning: techniques and applications

With impressive results across a wide range of use cases, gradient boosting machines (GBM) have quickly become a powerful tool in the field of cloud-based machine learning. The success of GBM, a strong ensemble learning approach, in managing complicated tasks while making use of cloud computing resources, has attracted a lot of interest. This study explores the methods and many uses of GBM in cloud settings. The equation above represents the core principle of gradient boosting, where each weak learner's contribution is optimized to improve the ensemble's performance iteratively. In (1) shows the gradient boosting algorithms, where $A_i(r)$ is the ensemble's prediction at iteration i for input x. $A_{i-1}(r)$ is the ensemble's prediction at iteration $i-1$ for input r. γ_i is the learning rate or step size for iteration i controlling the contribution of the weak learner $k_i(r)$ to the ensemble. $k_i(r)$ is the weak learner's prediction at iteration i for input r.

$$A_i(r) = A_{i-1}(r) + \gamma\ _i k_i(r) \tag{1}$$

Selecting an IDS based on its performance on a test dataset is typical practice. Two types of models are available for behaviour modelling: those based on supervised learning and those based on semi-supervised learning. The key distinction is whether or not the labels themselves are included in the provided training dataset. Intrusion detection models of several types are shown in Figure 1.

Figure 2 shows 2015–2020 GBM use. GBM, a strong ensemble learning method, is growing in popularity. Consumption has progressively increased from 320 units in 2015 to 1050 in 2020. This rise shows GBM's growing popularity in machine learning applications, particularly in cloud-based contexts where it provides scalable and accurate predictive modeling.

To generate a reliable prediction model, GBM repeatedly combines weak learners, often decision trees. Distributed computing resources are used most effectively with the combination of GBM and cloud computing. The training and prediction processes are sped up by using parallel processing methods to study data in chunks simultaneously. Distributed data storage is used by cloud-based GBM models to provide easy access and administration of massive datasets. Cloud-based automated model tuning makes use of grid search or Bayesian optimization to adjust hyper parameters. Credit scoring, fraud detection, and stock market prediction are just a few of the areas where cloud-based GBM models have excelled. Risk assessment models benefit from the cloud's capacity to analyze large financial information quickly and efficiently.

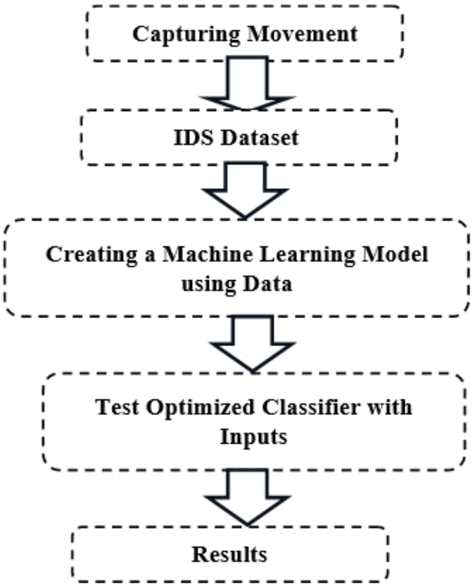

Figure 1. IDS block diagram for machine learning

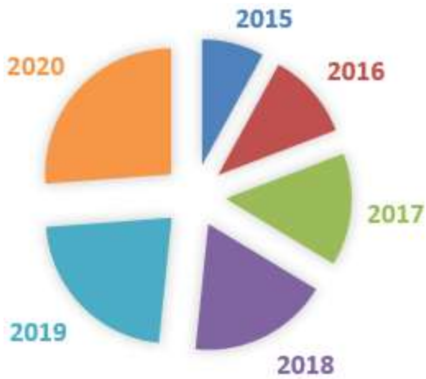

Figure 2. Yearwise consumption (2015-2020) of gradient boosting machines

2.2. Word2Vec in cloud-based machine learning: techniques and applications

Word2Vec is a game-changing NLP approach, and its seamless integration with cloud platforms means it can handle and interpret textual data like never before. The methods and many uses of Word2Vec in the cloud are investigated. By transforming words into dense vectors, Word2Vec is able to record the connections between words' meanings. Word2Vec's effectiveness and scalability have been improved by integration with cloud computing resources [27]. Word2Vec techniques using cloud computing can parallel process large text corpora, greatly accelerating model training time.

Access to large text datasets stored in the cloud is optimized, allowing for more straightforward training input. Word2Vec models trained on the cloud do very well at sentiment analysis, emotion recognition, and content classification. These algorithms can accurately anticipate sentiment by extracting detailed contextual information. When combined with cloud infrastructure, Word2Vec's ability to take semantic meaning into account enhances search engines. This improves the quality of the user's search results and suggestions. Word2Vec algorithms running in the cloud help recommendation systems learn about individual tastes from textual information. These algorithms provide customized suggestions by poring through customer feedback, product descriptions, and more. Cloud-based Word2Vec is a language translation model augmentation service [28]. Data collection, corpus generation, word embedding, DL model design and assessment, and experimental results are shown in Figure 3.

One of the popular algorithms within Word2Vec is skip-gram, and the equation for skip-gram can be summarized as follows: in (2) represents the Word2Vec algorithm where (w, c) represents a word–context pair from the training data. D is the set of all word-context pairs in the training data. $P(c|w)$ is the probability of observing the context word $c$given the target word w.

$$argmax \prod_{(w,c)} \in DP(c|w) \tag{2}$$

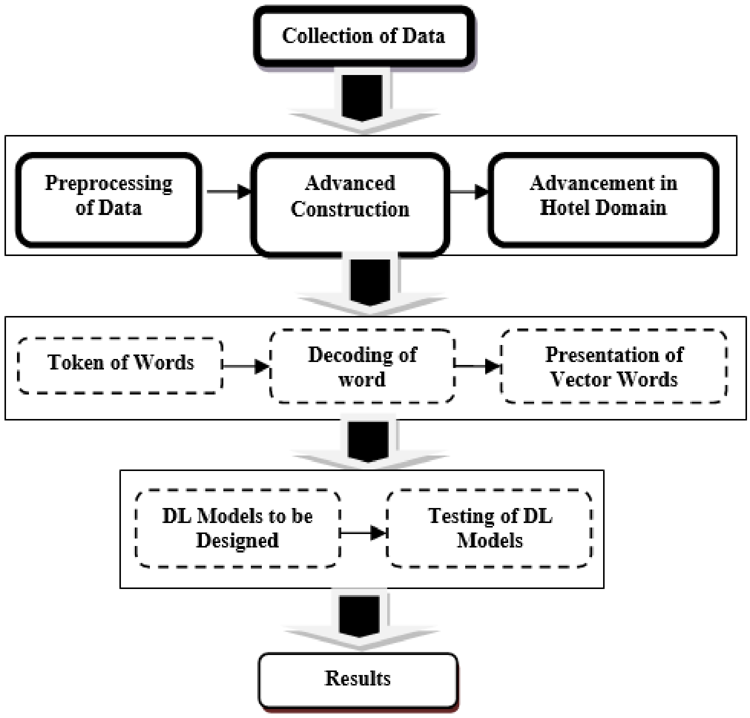

Figure 3. Combining Word2Vec with DL models for sentiment analysis

Table 1 presents a comparison of cloud-based machine learning, GBM, and Word2Vec with regards to their definitions, uses, benefits, and drawbacks. While GBM is an ensemble approach for robust prediction models, cloud-based machine learning makes use of cloud servers to make machine learning easily accessible and scalable. Word2Vec is an effective word embedding method for NLP. Because of their unique strengths and weaknesses, various situations call for different approaches.

Table 1. Comparison between cloud-based machine learning, GBM, and Word2Vec

Cloud-based ML	Gradient boosting machines	Word2Vec
Description	Scales and makes machine learning algorithms accessible on cloud servers.	Ensemble learning uses weak learners' outputs (typically trees) to create strong prediction models.
Application	Widely used for data analysis, pattern identification, and prediction.	Popular for classification and regression. Organized and unstructured data are handled.
Advantages	Flexible, remote, and cost-effective. Large dataset scalable.	Provides high predictive power, handles complex data, and reduces overfitting.
Challenges	Depends on network quality and safety. Privacy and compliance are hard.	Hyper parameter tweaking takes time. Overfitting occurs if not calibrated properly.

The process involves three parts. Annotating data comes first. Second, train the model, and finally, test it using cutting-edge research. The annotation process separates this data in two. The first 70% of data will train. The corpus will be evaluated using 30% of the second data set. Before training and testing feature extraction, each corpus undergoes various preprocessing procedures. The second section covers neural network multi-task models for noun phrase chunking. Figure 4 exhibits architectural system details.

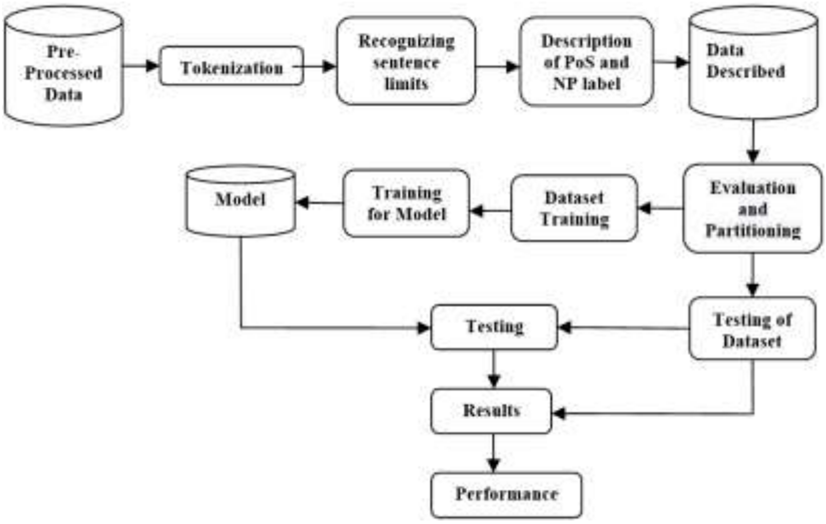

Figure 4. The system architecture of multi-task neural network for noun phrase chunking models

3. RESULTS AND DISCUSSION

3.1. Bayesian networks in cloud-based machine learning: techniques and applications

Combining probabilistic modeling with cloud computing, Bayesian networks in cloud-based machine learning enable complex data analysis and decision-making for a wide range of use cases [29], [30]. Cloud hosting is optimal for these networks because it allows the use of probabilistic graphical models to capture intricate dependencies between variables. The flexibility and availability of this connection are its greatest strengths, enabling the rapid processing of massive information and the coordinated efforts of users in different places. Bayesian networks in cloud-based machine learning enable organizations to make informed decisions, optimize processes, and navigate complex uncertainties with unparalleled precision and efficiency, whether in healthcare for diagnosis and treatment recommendations, finance for risk assessment, or industrial processes for predictive maintenance. New possibilities for data-driven insights and decision-making are made possible by the combination of Bayesian networks with cloud computing, which holds great promise for addressing a broad range of problems in a variety of fields of study and business.

Integrating Bayesian networks with cloud infrastructure improves risk assessment tools. They provide real-time analysis of data from several sources, allowing for more precise risk assessment and fraud detection. Natural catastrophes, disease outbreaks, and other urgent occurrences may be seen in the early stages with the help of Bayesian networks powered by the cloud, which analyzes environmental data. Their probabilistic basis makes them useful for making forecasts. These networks help with illness diagnosis, prognosis, and individualized therapy recommendations by combining patient data and medical information. The fundamental equation involves the joint probability distribution of variables A given the network structure E (3) shows the Bayesian networks where, $T(A|E)$ is the joint probability distribution of all variables A given the network structure E. A_k represents a random variable in the network. $Parents(A_k, E)$ are the parents of variable A_k in the DAG E. n is the total number of variables in the network.

$$T(A|E) = \Pi_{k=1}^{n} T(A_k|Parents(A_k, E)) \tag{3}$$

3.2. Gated recurrent units in cloud-based machine learning: techniques and applications

When combined with the huge processing capabilities of the cloud, gated recurrent units (GRU) in cloud-based machine learning revolutionize sequential data analysis. Because of its success in capturing long-range dependencies and mitigating the vanishing gradient issue, GRUs have become a popular kind of RNN. By taking use of cloud computing's scalability and parallel processing, GRUs are able to reach their full potential. NLP, audio identification, and time series forecasting are just some of the applications made possible by the rapid training and deployment of GRU models on enormous datasets. GRUs in cloud-based machine learning enable applications to process and analyze sequential data with unparalleled accuracy and efficiency, with potential applications ranging from virtual assistants and social media sentiment analysis to predictive maintenance in industrial settings. This synergy provides businesses with the means to draw useful conclusions, make sound choices, and provide superior customer experiences, ushering in a new age of data-driven offerings.

GRU's scalability and computing efficiency are both improved by its incorporation into the cloud. The use of models in areas requiring crucial decision-making will be aided by improved strategies for model interpretability. GRUs is a type of RNN architecture that aims to address the vanishing gradient problem and allow for the modeling of long-range dependencies in sequential data. The equations above describe the key components of a GRU's operation at a single time step t. In (4) to (6) shows the GRU, where j_i the concealed state at step i. s_i is the input at time step i. a_i is the update gate, calculating how much hidden state to update. p_i reset gate controls how much of the previous hidden state to forget. j'_i is the candidate hidden state, representing a new candidate for the hidden state. R_a, R_p are weight matrices for the gates and candidate hidden state. σ is the sigmoid activation function.

$$a_i = \sigma(R_a . [j_{i-1}, s_i]) \quad (4)$$

$$p_i = \sigma(R_p . [j_{i-1}, s_i]) \quad (5)$$

$$j'i = \tanh(Rj[pi[ji-1, si]]) \quad (6)$$

The online tool PHISHTANK was suggested for detecting phishing attempts. PHISHTANK uses many criteria to assess whether a website is safe or dangerous. Universal resource locator structure is developed to identify phishing attacks. The suggested research used machine learning methods and universal resource locator characteristics to tackle categorization challenges. Training characteristics were chosen based on an efficient phishing detection method. The suggested method architecture is shown in Figure 5.

Figure 6 shows 2015–2020 GRU use. Over the years, GRU, a recurrent neural network variation, has grown in use. To 550 units in 2020, consumption has progressively climbed from 280 in 2015. GRU sequential data modeling, particularly in cloud-based machine learning applications, allows efficient information flow and excellent predictions in voice recognition and time series forecasting, making it popular. Language modelling, sentiment analysis, and machine translation are just some of the NLP applications where cloud-based GRU models shine. They are able to better grasp and produce human language because of the context information they extract from text.

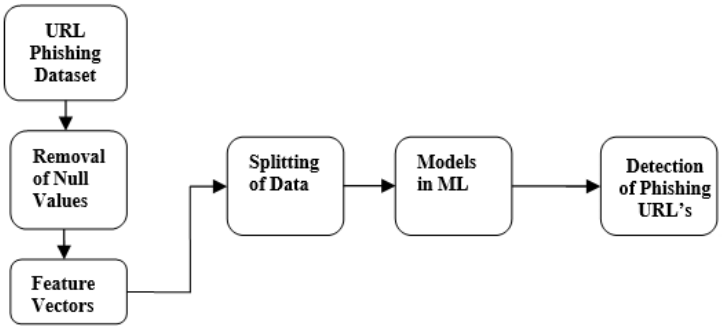

Figure 5. The general architecture of the PHISHTANK to detect a phishing attack using the URL

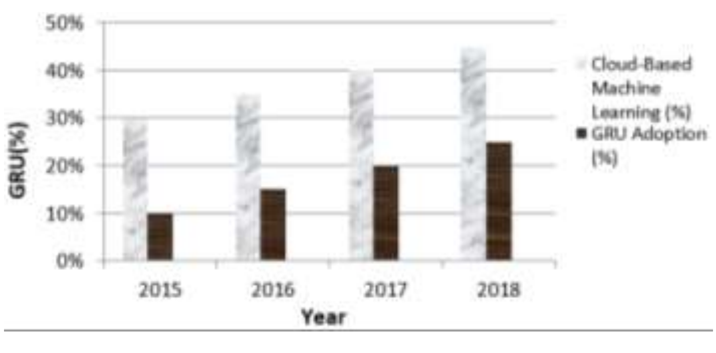

Figure 6. Year wise consumption (2015-2020) of gated recurrent units

4. CONCLUSION

GBM excels in predictive modeling, while Word2Vec aids in understanding semantic relationships in text. Bayesian networks offer probabilistic modeling in uncertain contexts and GRU excels in sequential data analysis. Leveraging cloud-based platforms enhances scalability. These techniques empower organizations to extract valuable insights from data, improve decision-making, and enhance operational efficiency. Bypassing such safeguards is an invasion into a computer network or system. Admitting the system's security breach is the most important step towards fixing it. The kind of contract, duration, monthly bill, and total cost most affect client turnover. Distributed edge/cloud apps, diverse networks, and fast-changing market needs make development harder. This complexity leads to hurriedly constructed and confusing software requirements, making it more fault-prone. The discipline has traditionally used machine learning, but deep learning is newer. Deep learning is replacing machine learning approaches in many industries, including business, owing to its capacity to increase accuracy with more training data, analyse complex and high-dimensional datasets, and need less human interaction.

REFERENCES

[1] M. Paricherla, M. Ritonga, S. R. Shinde, S. M. Chaudhari, R. Linur, and A. Raghuvanshi, "Machine learning techniques for accurate classification and detection of intrusions in computer network," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 4, pp. 2340–2347, Aug. 2023, doi: 10.11591/beei.v12i4.4708.

[2] S. Patil, M. Patil, and K. Chinnaiah, "Machine learning and deep learning for phishing page detection," *Research Reports on Computer Science*, pp. 45–54, May 2023, doi: 10.37256/rrcs.2320232629.

[3] B. K. Gontla, P. Gundu, P. J. Uppalapati, K. Narasimharao, and S. M. Hussain, "A Machine learning approach to identify phishing websites: a comparative study of classification models and ensemble learning techniques," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 5, pp. 1–9, Jun. 2023, doi: 10.4108/eetsis.vi.3300.

[4] T. A. R. Akbar and C. Apriono, "Machine learning predictive models analysis on telecommunications service churn rate," *Green Intelligent Systems and Applications*, vol. 3, no. 1, pp. 22–34, Jun. 2023, doi: 10.53623/gisa.v3i1.249.

[5] F. N. J. Muhamad, S. H. Ab Hamid, H. Subramaniam, R. Abdul Rashid, and F. Fahmi, "Fault-prone software requirements specification detection using ensemble learning for edge/cloud applications," *Applied Sciences (Switzerland)*, vol. 13, no. 14, p. 8368, Jul. 2023, doi: 10.3390/app13148368.

[6] M. Arifin, W. Widowati, and F. Farikhin, "Optimization of hyperparameters in machine learning for enhancing predictions of student academic performance," *Ingenierie des Systemes d'Information*, vol. 28, no. 3, pp. 575–582, Jun. 2023, doi: 10.18280/isi.280305.

[7] H. Attou, A. Guezzaz, S. Benkirane, M. Azrour, and Y. Farhaoui, "Cloud-based intrusion detection approach using machine learning techniques," *Big Data Mining and Analytics*, vol. 6, no. 3, pp. 311–320, Sep. 2023, doi: 10.26599/BDMA.2022.9020038.

[8] L. Saha, H. K. Tripathy, T. Gaber, H. El-Gohary, and E. S. M. El-kenawy, "Deep churn prediction method for telecommunication industry," *Sustainability (Switzerland)*, vol. 15, no. 5, p. 4543, Mar. 2023, doi: 10.3390/su15054543.

[9] A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari, and S. R. K. Joga, "Phishing detection system through hybrid machine learning based on URL," *IEEE Access*, vol. 11, pp. 36805–36822, 2023, doi: 10.1109/ACCESS.2023.3252366.

[10] S. Susanty, H. Sufriyana, E. C. Y. Su, and Y. H. Chuang, "Questionnaire-free machine-learning method to predict depressive symptoms among community-dwelling older adults," *PLoS ONE*, vol. 18, no. 1 January, p. e0280330, Jan. 2023, doi: 10.1371/journal.pone.0280330.

[11] R. Kusumaningrum, I. Z. Nisa, R. Jayanto, R. P. Nawangsari, and A. Wibowo, "Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews," *Heliyon*, vol. 9, no. 6, p. e17147, Jun. 2023, doi: 10.1016/j.heliyon.2023.e17147.

[12] Warto, Muljono, Purwanto, and E. Noersasongko, "Improving named entity recognition in Bahasa Indonesia with transformer-Word2Vec-CNN-attention model," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 4, pp. 655–668, Aug. 2023, doi: 10.22266/ijies2023.0831.53.

[13] S. R. Ningsih and A. Romadhony, "Difficulty level identification of Indonesian and mathematics multiple choice questions using machine learning approach," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3649.

[14] M. Alkaff, M. A. Miqdad, M. Fachrurrazi, M. N. Abdi, A. Z. Abidin, and R. Amalia, "Hate speech detection for Banjarese languages on Instagram using machine learning methods," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 3, pp. 495–504, Jul. 2023, doi: 10.30812/matrik.v22i3.2939.

[15] N. Khamphakdee and P. Seresangtakul, "An efficient deep learning for Thai sentiment analysis," *Data*, vol. 8, no. 5, p. 90, May 2023, doi: 10.3390/data8050090.

[16] A. Glenn, P. LaCasse, and B. Cox, "Emotion classification of Indonesian Tweets using bidirectional LSTM," *Neural Computing and Applications*, vol. 35, no. 13, pp. 9567–9578, Feb. 2023, doi: 10.1007/s00521-022-08186-1.

[17] J. Santoso, E. I. Setiawan, F. X. Ferdinandus, G. Gunawan, and L. Hernandez, "Indonesian language term extraction using multi-task neural network," *Knowledge Engineering and Data Science*, vol. 5, no. 2, p. 160, Dec. 2022, doi: 10.17977/um018v5i22022p160-167.

[18] F. Maulana and T. S. N. Asih, "Hoax classification in Indonesian language with bidirectional temporal convolutional network architecture," *Journal of Soft Computing Exploration*, vol. 4, no. 1, Jan. 2023, doi: 10.52465/joscex.v4i1.116.

[19] R. Rachidi, M. A. Ouassil, M. Errami, B. Cherradi, S. Hamida, and H. Silkan, "Classifying toxicity in the Arabic Moroccan dialect on Instagram: a machine and deep learning approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 1, pp. 588–598, Jul. 2023, doi: 10.11591/ijeecs.v31.i1.pp588-598.

[20] C. H. Chandra Sekhar, K. Vijayalakshmi, A. Sanyasi Rao, V. Vedanarayanan, M. B. Sahaai, and S. Murugan, "Cloud-based water tank management and control system," in *2023 2nd International Conference on Smart Technologies for Smart Nation, SmartTechCon 2023*, IEEE, Aug. 2023, pp. 641–646. doi: 10.1109/SmartTechCon57526.2023.10391730.

[21] I. Budi and R. R. Suryono, "Application of named entity recognition method for Indonesian datasets: a review," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 969–978, Apr. 2023, doi: 10.11591/eei.v12i2.4529.

[22] M. P. G. K. Jayaweera, W. M. C. J. T. Kithulwatta, and R. M. K. T. Rathnayaka, "Detect anomalies in cloud platforms by using network data: a review," *Cluster Computing*, vol. 26, no. 5, pp. 3279–3289, Jun. 2023, doi: 10.1007/s10586-023-04055-1.

[23] P. Rathnasiri, O. Adeniyi, and N. Thurairajah, "Data-driven approaches to built environment flood resilience: a scientometric and critical review," *Advanced Engineering Informatics*, vol. 57, p. 102085, Aug. 2023, doi: 10.1016/j.aei.2023.102085.

[24] K. T. Antwi-Agyakwa, M. K. Afenyo, and D. B. Angnuureng, "Know to predict, forecast to warn: a review of flood risk prediction tools," *Water (Switzerland)*, vol. 15, no. 3, p. 427, Jan. 2023, doi: 10.3390/w15030427.

[25] A. Aldelemy and Raed A. Abd-Alhameed, "Binary classification of customer's online purchasing behavior using machine learning," *Journal of Techniques*, vol. 5, no. 2, pp. 163–186, Jun. 2023, doi: 10.51173/jt.v5i2.1226.

[26] A. F. Sabbah and A. A. Hanani, "Self-admitted technical debt classification using natural language processing word embeddings," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 2142–2155, Apr. 2023, doi: 10.11591/ijece.v13i2.pp2142-2155.

[27] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/eei.v10i5.3157.

[28] C. S. Ranganathan, R. Raman, K. K. Sutaria, R. A Varma, and S. Murugan, "Network security in cyberspace using machine learning techniques," in *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Nov. 2024, pp. 1755–1759. doi: 10.1109/iceca58529.2023.10394962.

[29] M. J. Kumar, S. Mishra , G. R. Elangovan, M. Rajmohan , S. Murugan and N. A. Vignesh, "Bayesian decision model based reliable route formation in internet of things," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 3, pp. 1677-1685, 2024, doi: 10.11591/ijeecs.v34.i3.pp1665-1673.

[30] M. Amru *et al.*, "Network intrusion detection system by applying ensemble model for smart home," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 3, pp. 3485-3494, 2024, doi: 10.11591/ijece.v14i3.pp3485-3494.

BIOGRAPHIES OF AUTHORS

Raveendra N Amarnath is an accomplished Project Manager at Mphasis, boasting over 18 years of invaluable experience in the software industry. Residing in Dallas, Texas, his expertise shines in the realm of testing solutions, emphasizing reliability and automation. His career is characterized by his relentless pursuit of excellence in testing and his dedication to managing complex projects with precision. His extensive technical knowledge and leadership skills have made him an invaluable asset in the world of software and data management. In his current role, Raveendra spearheads the management of one of North America's most ambitious data migration projects, involving the seamless transition of data from Teradata to Google Cloud. His extensive background spans across diverse domains, including Airlines, Banking and Investment, Non-Life Insurance, Pharmacovigilance, and the UK Government. Raveendra's technical prowess encompasses data engineering, ETL (Extract, Transform, Load), informatica, teradata, big query, as well as mastery in JAVA and .NET applications. He is equally well-versed in the intricacies of SOAP and REST APIs. He can be contacted at email: raveendra.techie@gmail.com.

Gurumoorthi Gurulakshmanan is an experienced Technical Project Manager with 15+ years of experience leading multiple large-scale programs, developing strategic vision and goals, driving digital transformation to the cloud and evolving IT delivery methodologies. Proven record of championing solutions to solve highly-complex business problems and delivering on digital transformation with cloud-native solutions. Strong collaborator, partnered closely with key stakeholders to drive strategic approach to maximize growth. Influential and effective cross-functional team and people leader with a focus on cost optimization initiatives. A thought leader in Operational Efficiency and Team Building to ensure the best Customer Experience with a DevOps mindset. Demonstrated history of working in the industry as an Engineer, Technical Manager and Sr. Program Manager. He can be contacted at email: gurumoorthig198@gmail.com.