International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

REVIEW AND ANALYSIS ON REAL-WORLD APPLICATIONS OF SUPERVISED MACHINE LEARNING ALGORITHMS

V.Ramana Babu
Faculty in CSE dept KU, college of engineering and technology
Email id :ramana.vrb@gmail.com
R.Sushmitha
Faculty in CSE dept, KU college of engineering and technology
Email id: sushmacse511@gmail.com

ABSTRACT

Machine learning (ML) is a new and exciting area that opens up many doors for creative problem-solving in the real world. It's utilised in everything from fraud detection to reccomendation systems to medical imaging, and it allows machines to learn autonomously from data. There are three main types of ML, and they are supervised learning, unsupervised learning, and reinforcement learning. Supervised learning encompasses both classification and regression learning. Both methods require pre-training the model on a labelled dataset. When the output is continuous, regression can be applied. On the other hand, categorization is utilised for categorical outcomes. To this end, supervised learning employs predictor features to fine-tune class label models. Next, when the values of the predictor features are known but the value of the class label is unknown, a second classifier is employed to label the test data. During classification, each item in a training set is assigned a class based on its label. In regression, on the other hand, the label is a numeric response to the training data. There are a plethora of additional supervised learning strategies and algorithms out there, such as XGBoost, Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbor. In this overview paper, we take a look at supervised learning, discussing the pros and cons of various techniques and algorithms, performance indicators, and research findings.

1. INTRODUCTION

Upstream, midstream, and downstream all describe parts of the oil and gas sector. Reservoir characterization, drilling, and crude product production are all upstream processes. The output from the upstream is transferred to the midstream for further processing, storage, marketing, and transportation. Midstream outputs are collected, and then downstream operations like refining and distribution are carried out (PSAC, 2018). Machine learning has the potential to help solve a number of issues plaguing the oil and gas sector. It has been shown that machine learning may be used to predict pore pressure when drilling (Ahmed et al., 2019a), reservoir parameters in locations lacking core or suitable log data (Osarogiagbon et al., 2015), and oil production rate while characterising a reservoir (Ahmed et al., 2019b) (Mamudu et al., 2020).

Dangerous occurrences are unwelcome because they frequently result in squandered resources, diminished capacities, and even human lives. Some oil and gas accidents have clear origins that can be investigated and avoided in the future. As an extreme case, the Macondo blowout cost nearly $14 billion

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

(Mason, 2019). A kick, a fractured formation, a loss of circulation, and a clogged pipe are all potentially devastating results of drilling fluid density that should be avoided (Abimbola et al., 2015). Mishaps may occur if these occurrences are not watched and managed carefully. The multitude of variables that can affect the likelihood of these occurrences makes accurate forecasting difficult. Several cement-related issues (such as inadequate bonding and casing centralization) and pressure control equipment (such as anomalous pore pressure, insufficient mud density, and lost circulation) for controlled pressure drilling can cause kick (Tamim et al., 2019). This demonstrates the potential need for accurate mathematical models to anticipate or identify kick on the basis of contributing elements.

The success or failure of a machine learning solution will typically depend on the quality of the available data and the effectiveness of the underlying learning algorithms. Data-driven systems can be constructed with the help of a wide variety of machine learning algorithms. These include classification analysis, regression, data clustering, feature engineering and dimensionality reduction, association rule learning, and reinforcement learning. In addition, deep learning is a subset of the larger family of machine learning techniques, having evolved from the concept of the artificial neural network in the service of intelligent data analysis. This makes it challenging to identify a learning algorithm that will serve a given objective under certain conditions. This is due to the fact that, despite sharing some commonality, the outcomes produced by various learning algorithms for the same types of data might vary considerably. Numerous real-world applications (briefly described in Sect. "Applications of Machine Learning") rely on machine learning algorithms, making familiarity with their principles and their applicability crucial. These applications range from the Internet of Things to cybersecurity services to business and recommendation systems to smart cities to healthcare and COVID-19 to context-aware systems to sustainable agriculture.

This paper presents a detailed review of the numerous forms of machine learning algorithms that can be implemented to raise the intelligence and capacities of an application, since "Machine Learning" is a significant and potentially powerful tool for analysing the data described above. In light of this, the key contribution of this work is an elucidation of the principles and potential of various machine learning algorithms, as well as their applicability in the various real-world application domains previously outlined. The target audience for this paper consists of academics and developers from various academic and professional backgrounds who are interested in applying machine learning to the aforementioned disciplines in order to construct data-driven, automated, and intelligent systems.

2. LITERATURE REVIEW

The field of supervised learning has been the subject of many surveys during the past decade. For instance, the authors of paper [1] discuss the previous research on supervised learning-based classification techniques. In their study, they looked at five different types of classification systems and weighed their pros and cons (Naive Bayes, Neural Network, Decision tree, Support vector machine, and K-Nearest neighbour). They also sorted the papers by year of publication, classification system used, and research topic. Articles from the fields of medicine, agriculture, education, business, and networking are all included in their survey. Research shows that decision trees and Naive Bayes are the most popular

methods of classification. However, they were only able to survey five types of categorization systems.

The study's authors [2] classify supervised learning techniques into five groups: logic-based algorithms, statistical learning algorithms, instance-based learning, support vector machines, and deep learning. They also presented generic pseudocodes for a variety of learning algorithms, including decision trees, rule learners, Bayesian networks, and instance-based learners. Their research shows that neural networks and SVM perform better with continuous data. However, category data is where logic-based systems really shine. They also noted that Naive Bayes can perform adequately even with limited data. However, SVM and neural networks perform best when fed massive datasets. However, they didn't investigate regression in any detail, concentrating instead on classification algorithms.

Just as [3] summarises supervised classification methods, dividing them into those that rely on logic to learn, support vector machines, statistics, and laziness to learn, these methods are also discussed in detail. In this overview, the various methods are broken down and their advantages, disadvantages, and practical uses are discussed. The research concludes with a comparison of the efficacy of four widely-used algorithms utilising data from the Census Bureau Database (support vector machine, naive bayes, decision tree, and k-NN). They compared various systems based on several criteria, including classification speed, learning speed, and tolerance for background noise. With an accuracy of 84.94%, SVM was found to be superior to k-NN, Nave Bayes, and Decision Trees.

The researchers who compiled the report referred to as [4] examined the state of the art in supervised text categorization methods. NB, SVM, and k-NN were the machine learning methods surveyed, along with the metrics used to judge how well each performed. In addition, they discussed a number of different scoring systems for text analysis. Therefore, k-NN was the best performing ML algorithm. Based on the results of this research, it appears that algorithm performance varies depending on the dataset. Regrettably, only three types of classifiers were used in this investigation.

[5] compares several supervised learning methods based on their performance on real-world data. The eight comparison parameters used by the authors to evaluate the supervised learning algorithms' relative merits and shortcomings. Artificial Neural Networks, Logistic Regression, Naive Bayes, k-NN, Decision Trees, Random Forests, Bagged Trees, Memory-based Learning, and Boosted Stubs are all machine learning techniques (BS). Different metrics such as F-score, Cross entropy, ROC Curve, Squared error, Average Precision, Breakeven Point, and Accuracy, Precision, and Recall were considered. There was a general improvement in performance when calibrated enhanced trees were used. After SVM, Random Forest was a close third. However, logistic regression, Nave Bayes, and decision trees all fared poorly. The calibration of the models is surprisingly good, leading to very good results.

3. TYPES OF REAL-WORLD DATA AND MACHINE LEARNING TECHNIQUES

Data on people, companies, processes, transactions, events, and other entities is often gathered and analysed by machine learning algorithms. After discussing the general types of machine learning algorithms, we'll go on to discussing the many types of real-world data available.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

3.1 Types of Real-World Data

One of the most crucial aspects of creating a machine learning model or a data-driven practical solution is having access to appropriate amounts of data. Data can be in a variety of formats, including a more formalised structure, a less formal structure, or an entirely unstructured form. Another term for information about information is "metadata." In what follows, we'll discuss these details briefly.

- *Information is considered structured if it is arranged in accordance with a predetermined data model, can be readily retrieved, and is put to some kind of useful purpose, be it by a human or a machine. Structured data in well-defined schemes, such as relational databases, is typically stored in tabular form. Structured data includes things like names, dates, addresses, credit card information, stock prices, geolocation, etc.*
- *The lack of a uniform framework or organisation makes it challenging to collect, process, and analyse unstructured data, which consists primarily of text and multimedia. For example, sensor data, emails, blog posts, wiki pages, PDF files, audio files, video files, pictures, presentations, and web pages all fall under the category of unstructured data utilised in business.*
- *Although not stored in a relational database like the aforementioned structured data, semi-structured data still exhibits some of the same organisational qualities that aid in the comprehension of structured data. HTML, XML, JSON documents, NoSQL databases, etc. are only some of the common places to get information that is only loosely organised.*
- *Metadata is a special kind of data that is "data about data" rather than just data itself. To put it simply, data are anything that can be categorised, measured, or documented in relation to the data attributes of an organisation, while metadata are descriptions of those things. Metadata, on the other hand, provides a description of the important data information, elevating its value to data consumers. Metadata might include information about a document, such as who wrote it, how big the file is, when it was created, what it's about, and what it's used for.*

The fields of machine learning and data science make extensive use of a wide variety of standard datasets. Cybersecurity (NSL-KDD, UNSW-NB15, ISCX'12, CIC-DDoS2019, Bot-IoT, etc.) and smartphone data (phone call logs, SMS Log, mobile application usages logs, mobile phone notification logs, etc.); IoT data; agricultural and e-commerce data; health (heart disease, diabetes mellitus, COVID-19, etc.); and many other datasets. Depending on the needs of the specific real-world application, the data could be stored in any of the aforementioned formats. Learning capabilities and how various machine learning algorithms can be used to analyse such data in a specific problem area and extract insights or critical knowledge from the data for constructing real-world, intelligent applications are highlighted.

4. TYPES OF MACHINE LEARNING TECHNIQUES

Figure 1 shows that the majority of Machine Learning algorithms fall into one of four categories: supervised, unsupervised, semi-supervised, and reinforcement learning. Next,

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

we'll quickly review the fundamentals of each learning approach and how they may be applied to real-world problems.

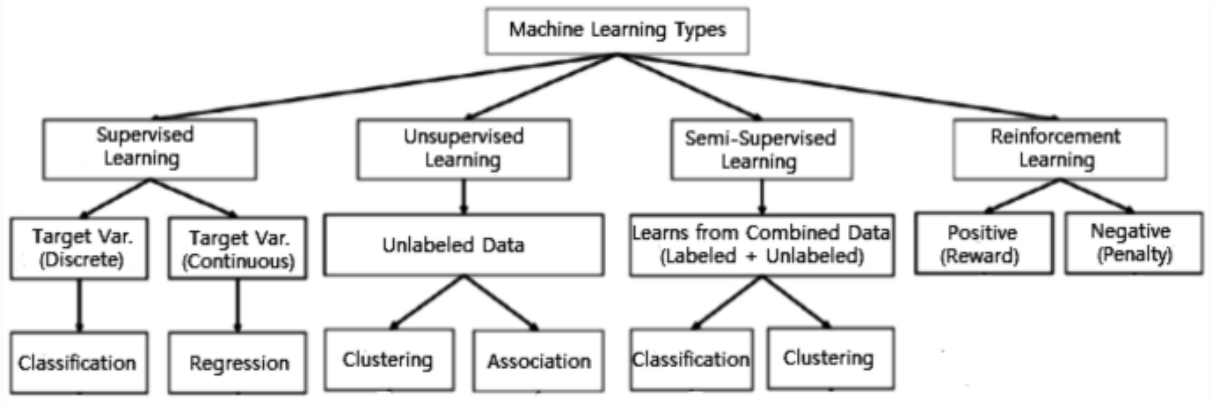

Fig. 1 The field of machine learning includes many different techniques..

- *Learning a function that maps inputs to outputs from a small set of examples has long been considered the holy grail of supervised machine learning. It learns a certain function by examining samples and training data that have been labelled. When predetermined outputs are required from a set of inputs, supervised learning is employed (i.e. a task-driven approach). Classification (data grouping) and regression are the two most common types of supervised jobs (data fitting). One use of supervised learning is text categorization, which involves making predictions about the category label or mood of a text such as a tweet or a review of a product.*
- *In contrast to supervised learning, unsupervised learning makes choices based only on the data without the assistance of human labelling. Common applications include the extraction of experimental motivations, noteworthy patterns and structures, clusters of results, and generative characteristics. Clustering, density estimation, feature learning, dimensionality reduction, creating association rules, anomaly detection, and many more are all examples of common unsupervised learning tasks.*
- *Semi-supervised learning closes the gap between the two types of learning by utilising both labelled and unlabeled input in its analysis. This style of learning bridges the gap between self-directed study and regular classroom instruction. In practise, semi-supervised learning is helpful since labelled data may be*

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

rare but unlabeled data may be copious. Semi-supervised learning models are developed with the intention of improving upon the prediction results obtained by using only the labelled data. Semi-supervised learning has found applications in many different areas, such as machine translation, fraud detection, data labelling, and text categorization.

- *Software agents and computers can use reinforcement learning algorithms to learn from their environments and decide for themselves what actions would produce the best outcomes. Ultimately, this type of education aims to help people use the knowledge they've received from environmental activists to maximise value or minimise risk.* Though it is not recommended for use in solving straightforward problems, it is a powerful tool for training AI models that can enhance the automation and operational efficiency of complex systems such as robots, self-driving vehicles, factories, and supply chains.

5. MACHINE LEARNING TASKS AND ALGORITHMS

Classification, regression, clustering, association rule learning, feature engineering for dimensionality reduction, and deep learning are only some of the machine learning applications covered. A machine learning-based predictive model, shown in a high-level overview in Fig. 2, is trained on historical data and then used to make predictions on fresh data.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

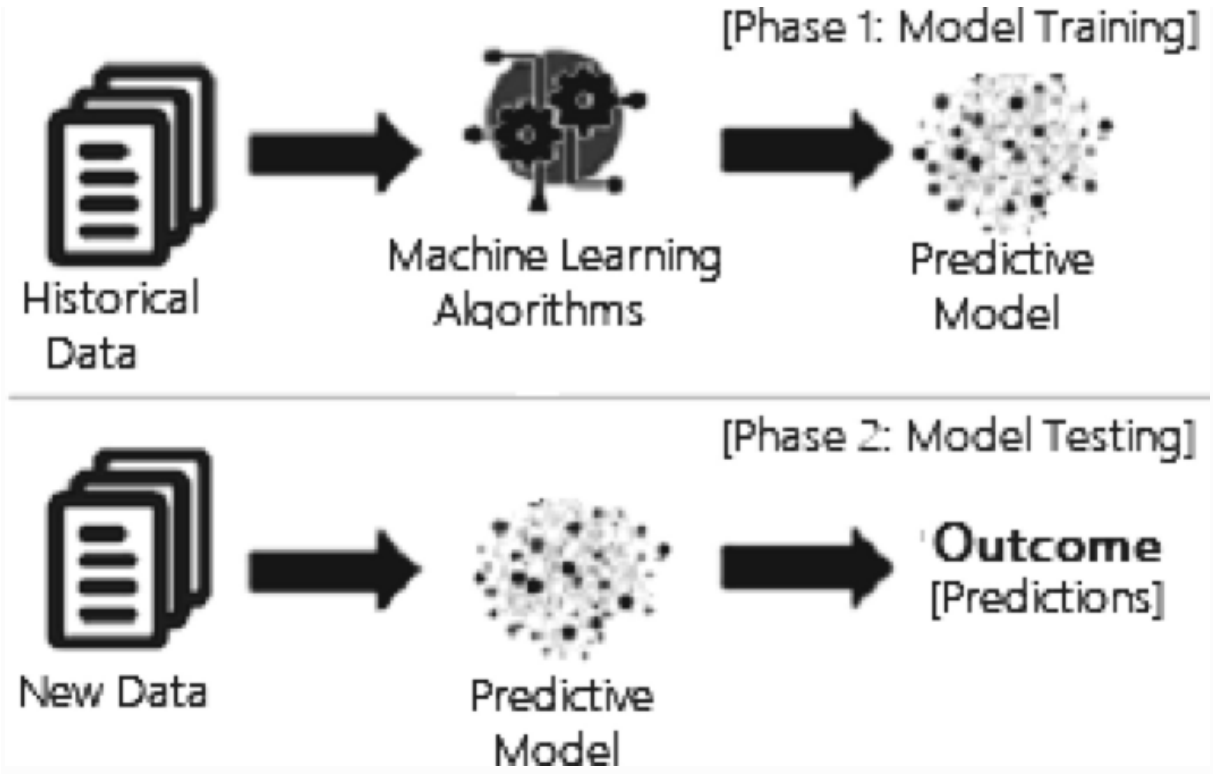

Fig. 2 An overarching framework for a machine learning-based predictive model that takes into account both the learning and validation stages

Classification Analysis

In the field of machine learning, classification refers to both a supervised learning method and a predictive modelling task in which a class label is predicted for an input example. Mathematically speaking, it converts X-variable inputs into Y-variable outputs (T, L, or C) (Y). It works on both structured and unstructured data and can be used to make predictions about the category of items being presented. For instance, email spam filtering can be thought of as a classification problem in which the inputs might be either spam or non spam. This section provides a high-level summary of the most common classification problems.

- Classification problems with only two possible answers (true or false) are known as "binary classification." For example, in a task requiring binary classification, "normal" could be one class and "abnormal" another. The "cancer not found" condition is the norm while doing an activity that necessitates a medical test, while the "cancer identified" state is the outlier. It is widely assumed that email providers use a similar binary approach to classifying messages as "spam" or "not spam."
- "Multiclass classification" is a common term for classification problems with more than two possible answers. Unlike binary classification problems, multiclass classification problems do not operate on the basis of normal and abnormal outcomes. Instead, each case is placed into one of several categories within that framework. The NSL-KDD dataset, for instance, classifies various network attacks as DoS (Denial of Service), U2R (User to Root), R2L (Root to

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Local), and Probing Attack. This means that spotting cyberattacks on a network is a complex multiclass classification problem.

- When a given example belongs to numerous classes, machine learning specialists take multi-label classification into account. Thus, multi-level text classification is an extension of multiclass classification where the classes involved are hierarchical and each example may simultaneously belong to more than one class at each hierarchical level. Google News, for instance, can be sorted and displayed in a number of different ways. In contrast to standard classification tasks, multi-label classification involves the use of sophisticated machine learning algorithms that allow for the prediction of multiple, independent classes or labels.

Numerous classification strategies have been proposed for use in machine learning and data science. We provide a high-level overview of the most often employed approaches in the following sections.

- Under the assumption of feature independence, the Bayes' theorem is implemented in the naive Bayes (NB) algorithm. It is useful for document or text classification, spam filtering, and many other real-world scenarios since it works well for binary and multi-class categories. The NB classifier can be used to efficiently and accurately identify noisy data samples, allowing for the construction of a reliable prediction model. In contrast to more complex algorithms, a minimal amount of training data is all that's required to make an accurate estimate of the required parameters. Strong assumptions it makes about the independence of qualities, however, may impede its performance. Many NB classifiers are based on distributions including the Gaussian, Multinomial, Complement, Bernays, and Categorical.
- By fitting conditional densities for each class to the data in accordance with Bayes' rule, statisticians are able to develop classifiers with linear decision boundaries. Linear discriminant analysis describes this approach (LDA). Projecting a dataset onto a lower-dimensional space, i.e. a drop in dimensionality that reduces the complexity of the model or the processing costs of running the model, is a key part of this method, which is a generalisation of Fisher's linear discriminant. Since it is expected that all classes share a common covariance matrix in a standard LDA model, a Gaussian distribution is a suitable fit for each class. Like ANOVA (analysis of variance) and regression analysis, LDA attempts to graphically express the relationship between two or more variables. In both cases, the dependent variable is modelled as a linear mixture of other features or measures.
- Regression Analysis: Logistic or Linear Logistic regression is another widely used probabilistic based statistical model for solving classification problems in machine learning (LR). Logistic regression often uses the sigmoid function (also known as the logistic function) given by Eq. 1 to estimate probabilities. Overfitting is prevented even with high-dimensional data, and the method really shines when the data can be neatly split in a linear form. Over-

fitting can be avoided with the use of regularisation (L1 and L2) methods. The assumption of linearity between the dependent and independent variables is a common criticism of logistic regression.

CONCLUSION

In this study, we provide a thorough evaluation of machine learning procedures for extracting insights from data and developing usable programmes. Our intention in this article was to provide a high-level summary of how machine learning techniques can be used to solve practical problems. The success of a machine learning model creation on on the accuracy of the input data and the efficiency of the applied learning algorithms. The complex learning algorithms need to be trained with the collected real-world data and information specific to the intended application before the system can give intelligent decision-making support. To further demonstrate the usefulness of machine learning approaches in addressing a wide range of practical problems, we also explored numerous common application domains. Finally, we have addressed and summed up the problems that have been encountered, as well as the prospects for future study and development. Therefore, effective answers to the mentioned issues are needed in a variety of application areas, creating promising research prospects in the subject.

REFERENCES

[1] James Cussens, "Machine Learning," IEEE Journal of Computing and Control, Vol.7, No.4, pp.164-168, 1996.

[2] Muhammad, I., & Yan, Z., "Supervised Machine Learning Approaches A Survey," ICTACT Journal on Soft Computing, Vol.5, No.3, 2015.

[3] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, Vol.31, No.3, pp.249-268, 2007.

[4] Richard S. Sutton and Andrew G. Barto, "Reinforcement Learning: An Introduction," Cambridge, MA: MIT Press, 1998.

[5] Sen, P. C., Hajra, M., & Ghosh, M., "Supervised classification algorithms in Machine Learning: A survey and review," Emerging technology in modelling and graphics, Springer, pp.99-111, 2020.

[6] Kadhim, A. I.,"Survey on supervised Machine Learning techniques for automatic text classification," AI Review, Vol.52, No.1, pp.273-292, 2019.

[7] Narayanan, U., Unnikrishnan, A., Paul, V., & Joseph, S., "A survey on various supervised classification algorithms," 2017 International Conference on Energy Communication, Data Analytics and Soft Computing (ICECDS), IEEE, pp.2118-2124, August 2017.

[8] Caruana, R., & Niculescu-Mizil, A., "An empirical comparison of supervised learning algorithms," Proceedings of the 23rd international conference on Machine Learning, pp.161-168, June 2006.

[9] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," Inf. Softw. Technol., Vol.51, No.1, pp.7-15, 2009.

[10] Tewari, A. S., Ansari, T. S., & Barman, A. G., "Opinion based book recommendation using naive bayes classifier," 2014 International Conference on Contemporary Computing and Informatics (IC3I), IEEE, pp.139-144, November 2014.

[11] Solanki, R. K., Verma, K., & Kumar, R., "Spam filtering using hybrid local-global Naive Bayes classifier," 2015 International Conference on Advances in Computing,

Communications and Informatics (ICACCI), IEEE, pp.829-833, August 2015.

[12] Jiang, Q., Wang, W., Han, X., Zhang, S., Wang, X., & Wang, C., "Deep feature weighting in Naive Bayes for Chinese text classification," 2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS), IEEE, pp.160-164, August 2016.

[13] Bhakre, S. K., & Bang, A., "Emotion recognition on the basis of audio signal using Naive Bayes classifier," 2016 International conference on advances in computing, communications and informatics (ICACCI), IEEE, pp.2363-2367, September 2016.

[14] Liu, J., Tian, Z., Liu, P., Jiang, J., & Li, Z., "An approach of semantic web service classification based on Naive Bayes," 2016 International Conference on Services Computing (SCC), IEEE, pp.356-362, June 2016.

[15] Liu, X., Lu, R., Ma, J., Chen, L., & Qin, B., "Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification," IEEE Journal of Biomedical and Health Informatics, Vol.20, No.2, pp.655-668, 2015.