ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 11 (2024) No. 1, pp. 181-194
DOI:10.33168/JLISS.2024.0112

Examining the Machine Learning Approaches for Identifying Significant Proteins in a Cancer Disease: A Systematic Literature Review

Angga Aditya Permana[1*], Ananda Setiyani Firman [1] , Analekta Tiara Perdana[2]

[1]Department of Informatics, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Banten, Indonesia

[2]Department of Biology, Faculty of Science, UIN Sultan Maulana Hasanuddin, Banten, Indonesia

angga.permana@umn.ac.id (Corresponding Author), ananda.setiyani@student.umn.ac.id, analekta.tiara@uinbanten.ac.id

Abstract. Previously, the medical field has struggled to identify suitable tools or systems to assist in researching and identifying significant proteins for specific diseases. However, recent advancements in artificial intelligence (AI) techniques, particularly in machine learning approaches, have led to researchers exploring this area. The primary objective of identifying significant proteins is to comprehend protein interactions and discover those that play pivotal roles in certain diseases especially cancer. By employing machine learning algorithms such as Principal Component Analysis (PCA), Support Vector Machine (SVM), and random forest, the healthcare sector can effectively work towards achieving these goals. In this systematic literature review, we collected, screened, and evaluated 42 articles published between 2013-2023, with a focus on identifying the most prominent machine learning algorithms utilized in this domain. The selection process involved considering articles with a Schimago Journal Rank score to ensure relevance and quality.

Keywords: Machine Learning Approach, PCA, Significant Protein, Study Literature Review

1. Introduction

In the medical field, doctors and researchers are trying to study diseases with the aim of finding suitable methods or treatments. Recently, treatment methods are divided into two: traditional medicine and modern medicine. Traditional medicine is the way knowledge, experience, and skills are passed down from generation to generation. While modern medicine is born out of scientific research. Personalized medicine is the strategies for utilizing an individual's distinct clinical, genomic, genetic, and environmental data to guide decisions about disease prevention, diagnosis, and treatment. Therapies administered to the patients were the best responses based upon on their individual features (Al-Tashi et al., 2023). In modern medicine, identifying protein significant associated with a disease is the first step in finding biomarker and drug designing so that it can be found the best way to treat the disease (Thakur et al., 2015). Protein significant identification is now mostly being performed using AI technology by studying and analyzing about the protein-protein interaction (PPI) network at genomic scale (Ochoa et al., 2015).

Physicians are just the same as other human being where they have an imperfection of human nature and sometimes are scarcity thus leading to misdiagnose in some terminal and serious disease (Akinnuwesi et al., 2020). Due to the costs and time-consuming of an experimental procedure in a few sets of proteins and a high-dimensional data already became easily accessible in all field of research, especially in medical field, they need some tools or systems that can help them to diagnose a terminal and serious disease with minimum time and cost (Constantino et al., 2020; Saha, Sengupta, et al., 2018). To satisfy the needs of computational systems, AI technique are helping medical team in diagnose a disease automatically and even give a better result (Ibrahim et al., 2021). Moreover, it can increase throughput and reduced the labor cost by using AI technique (Costantini et al., 2013). Machine learning approach is one of the AI techniques that can be used in drug discovery methodology, such as prediction target structure, prediction of biological activity of new ligands, and discovery or optimization hits (Chebouba et al., 2018).

This systematic review is specifically centered on examining machine learning approaches for identifying significant proteins in a cancer disease. Following its successful applications in cancer detection, the number of research works on machine learning-based approaches has increased exponentially recently (Nasser & Yusof, 2023). The motivation behind this research is the rapid growth in cancer incidence and mortality cases worldwide (Kumar et al., 2022). Scientists applied methods for screening in early stage, in order to find types of cancer before they cause symptoms. With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the medical research community. As a result, machine learning has become a popular tool for medical researchers (Kourou et al., 2015)

This begs for a systematic review and summary of the existing works to help successive researchers and practitioners gain better insight into and understanding of the field. Especially the contemporary problems in recognizing significant protein of a disease are of-ten based on a very complex and large volume of data. In the field of information technology, machine learning methods is playing a very important role such as helping fast development in both selection and extraction data (Topolski, 2020). As the machine learning used in selection and extraction data, it also aims to increase the efficiency of algorithms to combine classifiers into groups (Topolski & Topolska, 2019). One of the challenges that have been experienced in classification and data mining is classifying a high dimensional data set because it can be hard to visualize and understand (Lasisi & Attoh-Okine, 2018; Odhiambo Omuya et al., 2021). Data mining techniques, a method to identify the most related information in a dataset (Gewers et al., 2021), such as features extraction and reduction in data pre-processing process will improve a learning algorithm's performance significantly (Ed-daoudy & Maalmi, 2020). Moreover, machine learning algorithm can be used to de-sign multi-dimensional or image recognition even though it will consume a lot of data and memory (G. Zhang et al., 2019).

The used of machine learning approach in identifying significant protein will in-crease biomarker discovery. Finding significant protein by analyzing PPI network from the related disease can be used to find the biomarker or even drug designing (Karbalaei et al., 2018). Once the significant protein is found, biomarkers can be search based on the expression pat-terns that is revealed from the databases (Muhammad et al., 2019). Furthermore, it can also be used in drug repurposing of a new disease (Huot et al., 2021). There are some researches that has been done in creating a system or a new methodology

to predict, identify, or detect the significant protein of a disease such as WISCOD: a statistical web-enabled tool by (Vilardell et al., 2014) and a Multi Label Protein Function Prediction (ML_PFP) method by (Saha, Prasad, et al., 2018).

The purpose of this systematic literature review is to review research article in the past 10 years that studied about PPI and significant protein using machine learning approach. From the research review, it is expected that can be found the best machine learning approach to use in identifying significant protein of a cancer disease.

2. Literature Review

2.1. Systematic literature review

Systematic literature review is important for researcher, especially in health sciences and medical field, because it helps them in performing a new research by the previous result incrementally (Lame, 2019). Furthermore, seeing that it helps health sciences, systematic literature review has become the key methodology that keep refining the method to address new research questions.

2.2. Machine Learning

There are two types of machine learning, namely supervised and unsupervised machine learning. Supervised machine learning is machine learning algorithm that used to classify labelled datasets based on selected relevant features (Ed-daoudy & Maalmi, 2020; Odhiambo Omuya et al., 2021). Several research that used supervised machine learning are (Akinnuwesi et al., 2020; Ed-daoudy & Maalmi, 2020; Morais & Lima, 2018; Wang et al., 2018; Yin et al., 2022) using Support Vector Machine (SVM) and (Yin et al., 2022; Y. H. Zhang et al., 2021) using Random Forest (RF). While unsupervised machine learning is a method used to study the data structure, looking for the similarity of multiple objects, and to check the outliers in a dataset (Granato et al., 2018). Several research that used unsupervised machine learning are (Akinnuwesi et al., 2020; Arivudainambi et al., 2019; Ghufran et al., 2020; Granato et al., 2018; Ibrahim et al., 2021; Lim et al., 2019; Mahmoudi et al., 2021; Morais & Lima, 2018; Papi & Caracciolo, 2018; Wang et al., 2018; Wu et al., 2018) using Principal Component Analysis (PCA), (Ibrahim et al., 2021) using K-Nearest Neighbors (KNN), (Dickinson et al., 2018; Granato et al., 2018) using Hierarchal Clustering Algorithm (HCA), and (Liu et al., 2021; Wan et al., 2013) using ClusterONE. Further-more, there is deep learning approach that frequently learn a high level and sturdy at-tributes from the raw input and have been successfully applied to diverse classification and recognition task (Cao et al., 2018).

2.3. Bioinformatics

According to National Institutes of Health, bioinformatics can be defined as an application of tools that can compute and analyze biological data into visualization and commentation. The development of bioinformatics software, such as Qlucore Omics Explorer, has become an important part of disease therapy for targeting genetic and molecular treatments (Li et al., 2015). Other than that, there is Tissue Engineering (TE) that can provide the urge of tissue repair and/or regeneration of a human body (Beheshtizadeh et al., 2021).

2.4. Protein-Protein Interaction

Protein-Protein Interaction (PPI) is a sub-graphs with highly interconnected proteins that can be identified as protein complexes or functional modules (B. Kong et al., 2014). A protein complexes is formed by the interaction of two or more proteins (P. Kong et al., 2020). There are two challenging tasks in identifying the complexes of PPI network, first is a high through-put data interaction that has significantly high false positives and negatives. Second is a protein could belong to multiple complexes (Ramadan et al., 2016). In analyzing PPI network, centrality measurements are used as the parameters such as node degree, betweenness, and eigenvector (Malhotra et al., 2018; Vargas et al., 2016).

2.5. Related Work

Various machine learning approaches have been developed for identifying significant proteins in cancer disease. Al-Tashi et al. (2023) reviewed the machine learning models for the identification of prognostic and predictive cancer biomarkers. The results showed that the use of subgroup models as a means of predictive biomarker identification and suggested future directions for the field, including the integration of modern feature selection techniques such as metaheuristic methods and the enhancement of non-linear models by incorporating deep learning algorithms. Nasser and Yusof (2023) reviewed deep learning based methods for breast cancer diagnosis. The results showed that with the advent of AI, deep learning techniques have been used effectively in breast cancer detection, facilitating early diagnosis and therefore increasing the chances of patients' survival. It was required less human intervention for similar feature extraction. The Convolutional Neural Network (CNN) is the most accurate and extensively used model for breast cancer detection, and the accuracy metrics are the most popular method used for performance evaluation.

Kumar et al., (2022) reviewed AI technique in cancer prediction and diagnosis. The results showed that deep learning and machine learning models provide a reliable, rapid, and effective solution. Although multiple techniques recommended in the literature have achieved great prediction results, still cancer mortality has not been reduced. Thus, more extensive research to deal with the challenges of cancer prediction is required. Kourou et al., (2015) reviewed machine learning applications in cancer prognosis and prediction. A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making.

Cruz & Wishart, (2006) reviewed applications of machine learning in cancer prediction and prognosis growing trend towards personalized, predictive medicine. Broad survey was conducted about the different types of machine learning methods being used, the types of data being integrated and the performance of these methods in cancer prediction and prognosis. Artificial neural networks (ANNs) was the most usable machine learning method. Nassif et al., (2022) reviewed breast cancer detection using artificial intelligence techniques. Artificial intelligence and machine learning have been used effectively in detection and treatment of several dangerous diseases, helping in early diagnosis and treatment, and thus increasing the patient's chance of survival. Deep learning has been designed to analyze the most important features affecting detection and treatment of serious diseases. CNN algorithm was the most widely used for both gene expression and MRI images data, because of its good results in comparison to other algorithms.

3. Research Method

This systematic review was conducted in order to gain literature information about the usage of machine learning in reducing feature set to find classification of significant protein. This study employed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guiding principles for conducting systematic reviews.

3.1. Research question

1. RQ-1: Which journal publish article about research of significant protein using machine learning method?
2. RQ-2: Which type of machine learning method is used in significant protein analysis?
3. RQ-3: Which type of algorithm that used in machine learning method for significant protein analysis?
4. RQ-4: Which type of algorithm that used in machine learning method for finding biomarker in a disease?

Based on the research questions above, furthermore will be used to decide the search keywords in the next step.

3.2. Record identification and screening

3.2.1 Keyword

In order to obtain a comprehensive search string, the keywords used in this re-search will be based on the research questions term, basic and affix word according to the research theme, and usage of "AND" and "OR" adjust with the needs. Information sources were search in Publish or Perish (PoP) application using keywords such as:

- *"Significant Protein"*
- *"Leukemia" AND "Significant Protein"*
- *"Significant Protein" AND "Machine Learning"*
- *"Principal Component Analysis" AND "Significant Protein"*
- *"Cancer" AND "Significant Protein"*
- *"ClusterONE"*
- *"Principal Component Analysis" AND "Cancer"*

The search strings above were used as search keywords for the articles that used in the research. Each of the articles were filtered by articles indexes by scopus between Quartile 1 until Quartile 4. For each keyword, obtained 7 to 200 articles that related to the research topic but only 42 articles that meet the criteria of this research. In Figure 1 can be seen the distribution of articles that meet the criteria with each of their quartile.

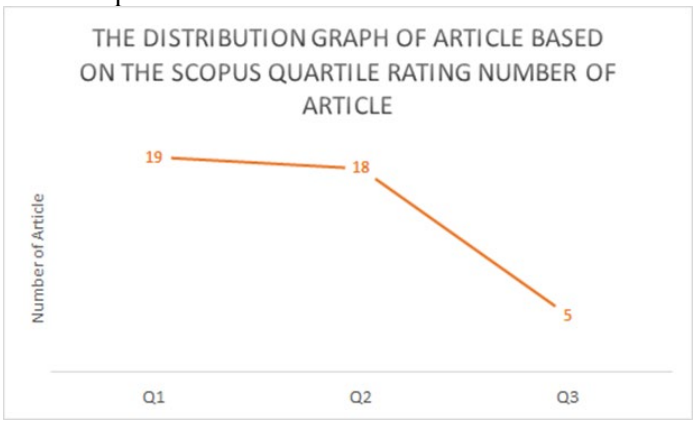

Fig.1: The distribution graph of article based on the SQR

3.2.2 Quality evaluation

1. Inclusion and exclusion criteria

Table 1. Inclusion and exclusion criteria

No	Inclusion Criteria	Exclusion Criteria
1	Research articles published between 2013-2023	Research articles published before 2013
2	Research articles published in English	Research articles published in Bahasa
3	Research journals indexed by Scopus	Research journals not indexed by Scopus
4	Research articles discussing the use of machine learning in	Research articles discussing significant protein analysis without using machine learning

	significant protein analysis	

2. PRISMA Protocol

Once inclusion criteria were decided and used on filtering the research articles, there is protocol called Preferred Reporting Items for Systematic Literature Review and Meta-analysis (PRISMA) that used to eliminate research articles. These are the steps to describe the following protocol:

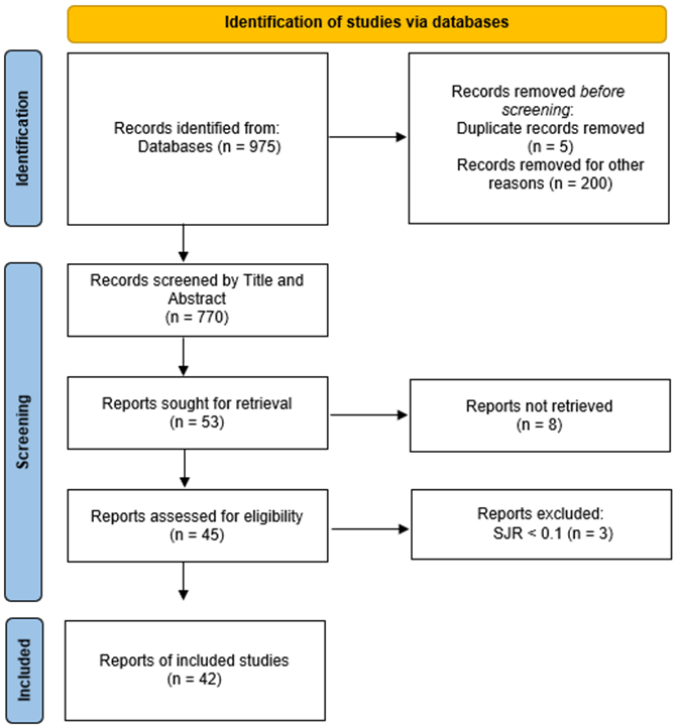

Fig.2: PRISMA flow diagram

4. Result and Discussion

4.1. Research article obtained by keyword search

The number of publications regarding machine learning approaches for identifying significant proteins in a cancer disease from 2018 until 2023 is illustrated in Table 2, with the total 932 articles. The keywords that were created based on the research question were used in the PoP application to find the research articles that related to the research topic.

Table 2. Number of research articles based on keywords

No	Keywords	Inclusion Criteria
1	“Significant Protein”	417 articles
2	“Leukemia” AND “Significant Protein”	7 articles
3	“Significant Protein” AND “Machine Learning”	48 articles
4	“PCA” AND “Significant Protein”	150 articles
5	“Cancer” AND “Significant Protein”	194 articles
6	“ClusterONE”	116 articles

4.2. Research article based on keywords and scopus quartile

From the total of research articles collected based on the keyword search, there were 42 articles in total that were related to the research topic and had Score Journal Ranking (SJR) score. In Figure 3 is a graph to visualize the distribution of articles that meet the requirements based on the SJR score. Green dots

represent the article that has Quartile 1 with 19 articles, yellow dots represent the article that has Quartile 2 with 18 articles, and orange dots represent the article that has Quartile 3 with 5 articles.

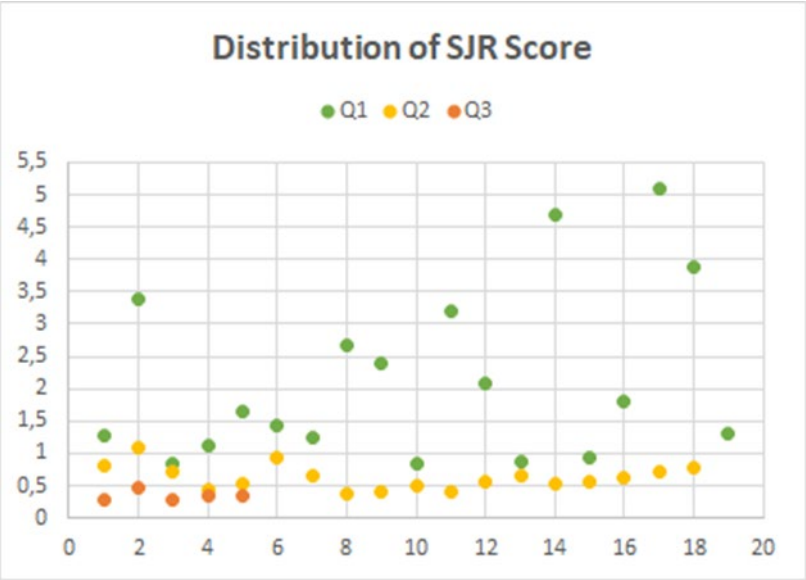

Fig.3: Graph of distribution of SJR score

Figure 4 is a graph to visualize the topic and objective of overall research articles that were collected using the keywords before. From the articles, the top 5 machine learning algorithms are PCA, SVM, RF, Naïve Bayes and KNN. Moreover, the top 5 diseases are cancer, leukemia, asthma, type II diabetes and alzheimer.

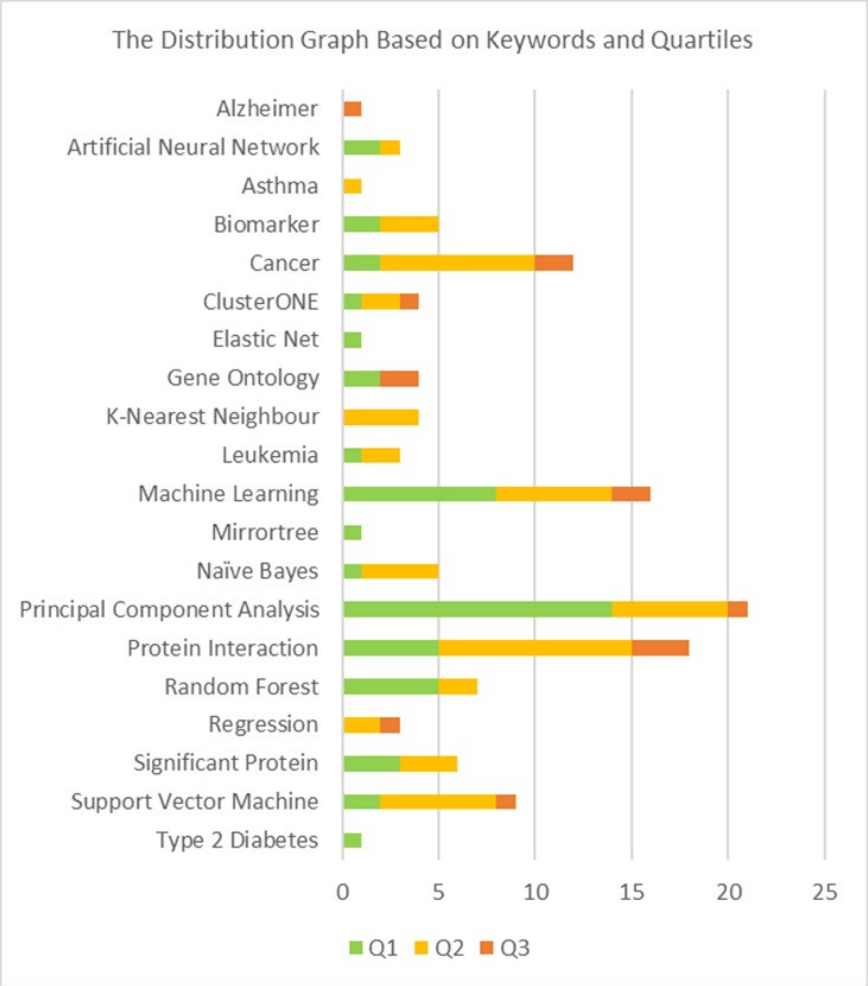

Fig.4: Graph of distribution article based on keywords and quartiles

4.3. Machine learning algorithms

From the articles that were collected and being eliminated based on the criteria before, machine learning algorithms was one of the main points to be considerate. The machine learning algorithms that were used in the research articles can become the object of comparison for this systematic literature review. In Figure 5 it can be seen the machine learning algorithms that used in the articles that were collected and compared to the quartile of the articles. The most machine learning method used in the research articles that were collected is PCA and has Quartile 1 at most.

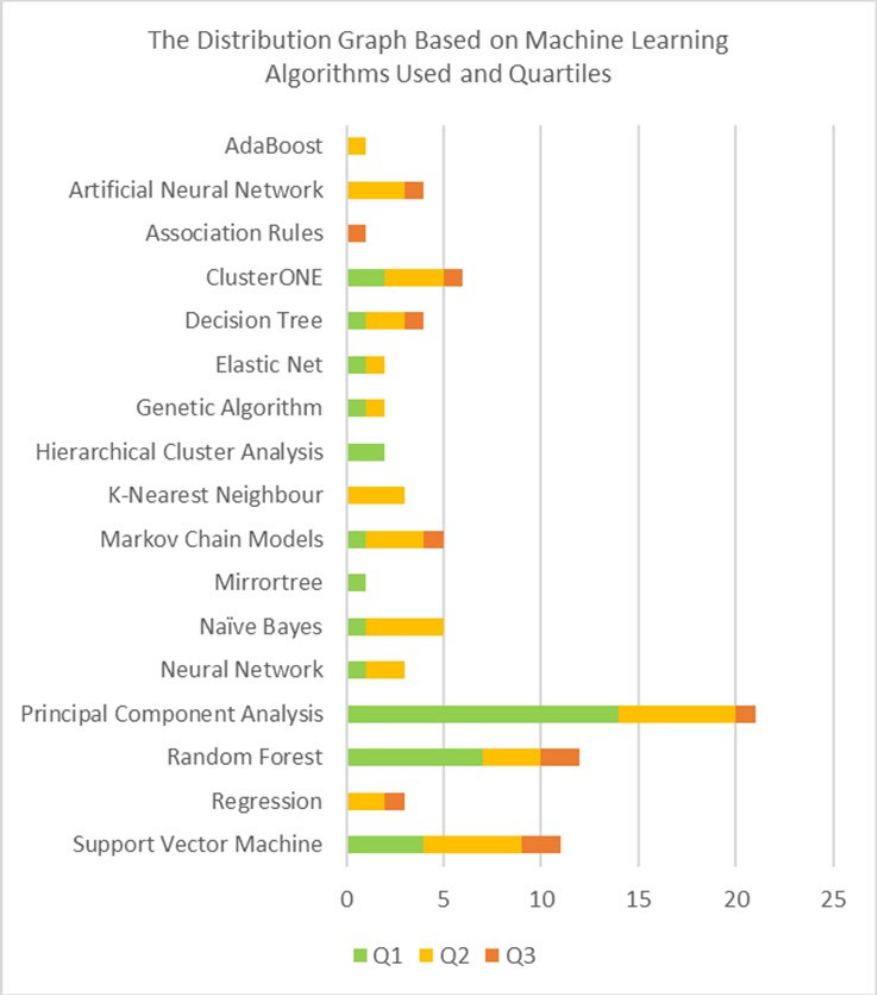

Fig.5: Graph of distribution articles based on machine learning algorithms

PCA is a statistical method to analyze multivariate data by reducing the number of variables for data analysis and interpretation. It reduces dimensionality at once retains the variance of the multivariate data. PCA has been used to analyzed the differential expression of proteins in breast cancer with the percentage of correct classification was 91.7% for the originally grouped tissue samples and 88.9% for cross-validated samples (Liang et al., 2010). Elhaik (2022)demonstrated that PCA can generate desired outcomes but still needs an alternative mixed-admixture population genetic model. Sell et al., (2020) explored potential disease identification in high dimensional blood microRNA (miRNA) datasets using PCA. PCA was proved as identifier of patients with specific disease, such as heart disease, stroke, hypertension, sepsis, diabetes, cancer HIV hemophilia, meningitis, multiple sclerosis, and so on.

4.4. Machine learning types

Classification is the most used type of machine learning for articles related to machine learning approaches for precision medicine, followed by regression and ensemble types of machine learning. In Figure 6 it can be seen the visualization of machine learning types distribution of the articles that is used. From all algorithms: SVM, RF, Naive Bayes, KNN and Decision Tree are included into classification type. The rest of the algorithms are included into clustering, regression and dimensionality reduction. Classification algorithms would be effective in discriminating ovarian cancer from benign disease and healthy controls. It was also potential tool for the analysis of high-dimensional proteomic data (Vlahou et al., 2003).

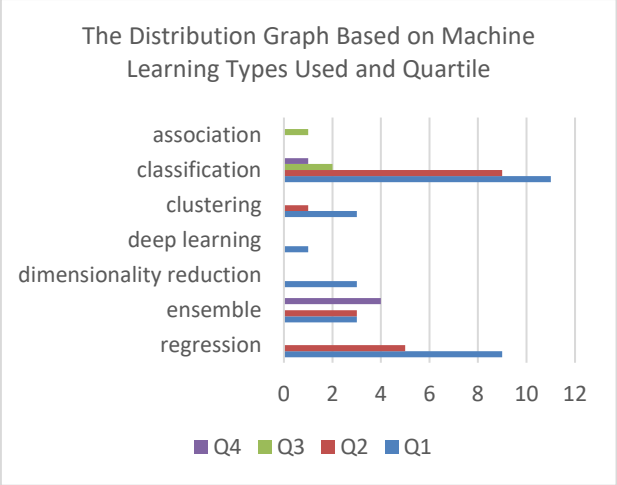

Fig. 6: The distribution graph of machine learning types

5. Conclusion

This paper of a systematic literature review, covering the articles that mainly discuss about identifying significant protein of a disease especially cancer using machine learning approach as the algorithms. Our study concluded that most previous literature works employed PCA. Another significant factor noted is that most studies is about cancer. Also, the classification algorithms are the intend to use. Although multiple pieces of research have displayed, there is still a need to address the challenges of machine learning approaches for identifying significant proteins in a cancer disease.

Acknowledgment

The authors would like to thank Universitas Multimedia Nusantara for the support of this research work.

References

Akinnuwesi, B. A., Macaulay, B. O., & Aribisala, B. S. (2020). Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques. *Informatics in Medicine Unlocked*, *21*. https://doi.org/10.1016/j.imu.2020.100459

Al-Tashi, Q., Saad, M. B., Muneer, A., Qureshi, R., Mirjalili, S., Sheshadri, A., Le, X., Vokes, N. I., Zhang, J., & Wu, J. (2023). Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic Review. *International Journal of Molecular Sciences*, *24*(9). https://doi.org/10.3390/ijms24097781

Arivudainambi, D., Varun, V. K., S., S. C., & Visu, P. (2019). Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance. *Computer Communications*, *147*, 50–57. https://doi.org/10.1016/j.comcom.2019.08.003

Beheshtizadeh, N., Asgari, Y., Nasiri, N., Farzin, A., Ghorbani, M., Lotfibakhshaiesh, N., & Azami, M. (2021). A network analysis of angiogenesis/osteogenesis-related growth factors in bone tissue engineering based on in-vitro and in-vivo data: A systems biology approach. *Tissue and Cell*, *72*. https://doi.org/10.1016/j.tice.2021.101553

Cao, W., Czarnek, N., Shan, J., & Li, L. (2018). Microaneurysm detection using principal component analysis and machine learning methods. *IEEE Transactions on Nanobioscience*, *17*(3), 191–198. https://doi.org/10.1109/TNB.2018.2840084

Chebouba, L., Boughaci, D., & Guziolowski, C. (2018). Proteomics Versus Clinical Data and Stochastic Local Search Based Feature Selection for Acute Myeloid Leukemia Patients' Classification. *Journal of*

Medical Systems, *42*(7). https://doi.org/10.1007/s10916-018-0972-z

Constantino, C., Carvalho, A. M., & Vinga, S. (2020). Sparse Consensus Classification for Discovering Novel Biomarkers in Rheumatoid Arthritis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12565 LNCS*, 125–136. https://doi.org/10.1007/978-3-030-64583-0_13

Costantini, S., Capone, F., Maio, P., Guerriero, E., Colonna, G., Izzo, F., & Castello, G. (2013). Cancer biomarker profiling in patients with chronic hepatitis C virus, liver cirrhosis and hepatocellular carcinoma. *Oncology Reports*, *29*(6), 2163–2168. https://doi.org/10.3892/or.2013.2378

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, *2*, 59–77. https://doi.org/10.1177/117693510600200030

Dickinson, A., Saraswat, M., Mäkitie, A., Silén, R., Hagström, J., Haglund, C., Joenväärä, S., & Silén, S. (2018). Label-free tissue proteomics can classify oral squamous cell carcinoma from healthy tissue in a stage-specific manner. *Oral Oncology*, *86*, 206–215. https://doi.org/10.1016/j.oraloncology.2018.09.013

Ed-daoudy, A., & Maalmi, K. (2020). Breast cancer classification with reduced feature set using association rules and support vector machine. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *9*(1). https://doi.org/10.1007/s13721-020-00237-8

Elhaik, E. (2022). Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. In *Scientific Reports* (Vol. 12, Issue 1). Nature Publishing Group UK. https://doi.org/10.1038/s41598-022-14395-4

Gewers, F. L., Ferreira, G. R., De Arruda, H. F., Silva, F. N., Comin, C. H., Amancio, D. R., & Costa, L. D. F. (2021). Principal component analysis: A natural approach to data exploration. *ACM Computing Surveys*, *54*(4). https://doi.org/10.1145/3447755

Ghufran, M., Rehman, A. U., Shah, M., Ayaz, M., Ng, H. L., & Wadood, A. (2020). In-silico design of peptide inhibitors of K-Ras target in cancer disease. *Journal of Biomolecular Structure and Dynamics*, *38*(18), 5488–5499. https://doi.org/10.1080/07391102.2019.1704880

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science and Technology*, *72*, 83–90. https://doi.org/10.1016/j.tifs.2017.12.006

Huot, M., Caron, M., Richer, C., Djibo, R., Najmanovich, R., St-Onge, P., Sinnett, D., & Raynal, N. J. M. (2021). Repurposing proscillaridin A in combination with decitabine against embryonal rhabdomyosarcoma RD cells. *Cancer Chemotherapy and Pharmacology*, *88*(5), 845–856. https://doi.org/10.1007/s00280-021-04339-6

Ibrahim, S., Nazir, S., & Velastin, S. A. (2021). Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. *Journal of Imaging*, *7*(11). https://doi.org/10.3390/jimaging7110225

Karbalaei, R., Allahyari, M., Rezaei-Tavirani, M., Asadzadeh-Aghdaei, H., & Zali, M. R. (2018). Protein-protein interaction analysis of alzheimer's disease and NAFLD based on systems biology methods unhide common ancestor pathways. *Gastroenterology and Hepatology from Bed to Bench*, *11*(1), 27–33. https://doi.org/10.22037/ghfbb.v0i0.1327

Kong, B., Yang, T., Chen, L., Kuang, Y. Q., Gu, J. W., Xia, X., Cheng, L., & Zhang, J. H. (2014). Protein-protein interaction network analysis and gene set enrichment analysis in epilepsy patients with brain cancer. *Journal of Clinical Neuroscience*, *21*(2), 316–319. https://doi.org/10.1016/j.jocn.2013.06.026

Kong, P., Huang, G., & Liu, W. (2020). Identification of protein complexes and functional modules in E. coli PPI networks. *BMC Microbiology*, *20*(1). https://doi.org/10.1186/s12866-020-01904-6

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Kumar, Y., Gupta, S., Singla, R., & Hu, Y. C. (2022). A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis. *Archives of Computational Methods in Engineering*, *29*(4), 2043–2070. https://doi.org/10.1007/s11831-021-09648-w

Lame, G. (2019). Systematic Literature Reviews: An Introduction. *Proceedings of the Design Society: International Conference on Engineering Design*, *1*(1), 1633–1642. https://doi.org/10.1017/dsi.2019.169

Lasisi, A., & Attoh-Okine, N. (2018). Principal components analysis and track quality index: A machine learning approach. *Transportation Research Part C: Emerging Technologies*, *91*, 230–248. https://doi.org/10.1016/j.trc.2018.04.001

Li, Z., Qiao, Z., Zheng, W., & Ma, W. (2015). Network Cluster Analysis of Protein-Protein Interaction Network-Identified Biomarker for Type 2 Diabetes. *Diabetes Technology & Therapeutics*, *17*(7), 475–481. https://doi.org/10.1089/dia.2014.0204

Liang, S., Singh, M., Dharmaraj, S., & Gam, L. H. (2010). The PCA and LDA analysis on the differential expression of proteins in breast cancer. *Disease Markers*, *29*(5), 231–242. https://doi.org/10.3233/DMA-2010-0753

Lim, J. Y., Nam, J. S., Shin, H., Park, J., Song, H. I., Kang, M., Lim, K. Il, & Choi, Y. (2019). Identification of Newly Emerging Influenza Viruses by Detecting the Virally Infected Cells Based on Surface Enhanced Raman Spectroscopy and Principal Component Analysis. *Analytical Chemistry*, *91*(9), 5677–5684. https://doi.org/10.1021/acs.analchem.8b05533

Liu, G., Liu, B., Li, A., Wang, X., Yu, J., & Zhou, X. (2021). Identifying Protein Complexes With Clear Module Structure Using Pairwise Constraints in Protein Interaction Networks. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.664786

Mahmoudi, M. R., Heydari, M. H., Qasem, S. N., Mosavi, A., & Band, S. S. (2021). Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries. *Alexandria Engineering Journal*, *60*(1), 457–464. https://doi.org/10.1016/j.aej.2020.09.013

Malhotra, A. G., Jha, M., Singh, S., & Pandey, K. M. (2018). Construction of a Comprehensive Protein–Protein Interaction Map for Vitiligo Disease to Identify Key Regulatory Elements: A Systemic Approach. *Interdisciplinary Sciences – Computational Life Sciences*, *10*(3), 500–514. https://doi.org/10.1007/s12539-017-0213-z

Morais, C. L. M., & Lima, K. M. G. (2018). Principal component analysis with linear and quadratic discriminant analysis for identification of cancer samples based on mass spectrometry. *Journal of the Brazilian Chemical Society*, *29*(3), 472–481. https://doi.org/10.21577/0103-5053.20170159

Muhammad, S. A., Fatima, N., Paracha, R. Z., Ali, A., & Chen, J. Y. (2019). A systematic simulation-based meta-analytical framework for prediction of physiological biomarkers in alopecia. *Journal of Biological Research (Greece)*, *26*(1). https://doi.org/10.1186/s40709-019-0094-x

Nasser, M., & Yusof, U. K. (2023). Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. *Diagnostics*, *13*(1). https://doi.org/10.3390/diagnostics13010161

Nassif, A. B., Talib, M. A., Nasir, Q., Afadar, Y., & Elgendy, O. (2022). Breast cancer detection using

artificial intelligence techniques: A systematic literature review. *Artificial Intelligence in Medicine*, *127*(May). https://doi.org/10.1016/j.artmed.2022.102276

Ochoa, D., Juan, D., Valencia, A., & Pazos, F. (2015). Detection of significant protein coevolution. *Bioinformatics*, *31*(13), 2166–2173. https://doi.org/10.1093/bioinformatics/btv102

Odhiambo Omuya, E., Onyango Okeyo, G., & Waema Kimwele, M. (2021). Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Systems with Applications*, *174*. https://doi.org/10.1016/j.eswa.2021.114765

Papi, M., & Caracciolo, G. (2018). Principal component analysis of personalized biomolecular corona data for early disease detection. *Nano Today*, *21*, 14–17. https://doi.org/10.1016/j.nantod.2018.03.001

Ramadan, E., Naef, A., & Ahmed, M. (2016). Protein complexes predictions within protein interaction networks using genetic algorithms. *BMC Bioinformatics*, *17*. https://doi.org/10.1186/s12859-016-1096-4

Saha, S., Prasad, A., Chatterjee, P., Basu, S., & Nasipuri, M. (2018). Protein function prediction from protein-protein interaction network using gene ontology based neighborhood analysis and physico-chemical features. *Journal of Bioinformatics and Computational Biology*, *16*(6). https://doi.org/10.1142/S0219720018500257

Saha, S., Sengupta, K., Chatterjee, P., Basu, S., & Nasipuri, M. (2018). Analysis of protein targets in pathogen-host interaction in infectious diseases: A case study on Plasmodium falciparum and Homo sapiens interaction network. *Briefings in Functional Genomics*, *17*(6), 441–450. https://doi.org/10.1093/bfgp/elx024

Sell, S. L., Widen, S. G., Prough, D. S., & Hellmich, H. L. (2020). Principal component analysis of blood microRNA datasets facilitates diagnosis of diverse diseases. *PLoS ONE*, *15*(6 June), 1–26. https://doi.org/10.1371/journal.pone.0234185

Thakur, S., Dhiman, M., Tell, G., & Mantha, A. K. (2015). A review on protein-protein interaction network of APE1/Ref-1 and its associated biological functions. *Cell Biochemistry and Function*, *33*(3), 101–112. https://doi.org/10.1002/cbf.3100

Topolski, M. (2020). The modified principal component analysis feature extraction method for the task of diagnosing chronic lymphocytic leukemia type b-cll. *Journal of Universal Computer Science*, *26*(6), 734–746. https://doi.org/10.3897/jucs.2020.039

Topolski, M., & Topolska, K. (2019). Algorithm for Constructing a Classifier Team Using a Modified PCA (Principal Component Analysis) in the Task of Diagnosis of Acute Lymphocytic Leukaemia Type B-CLL. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11734 LNAI*, 614–624. https://doi.org/10.1007/978-3-030-29859-3_52

Vargas, J. E., Porto, B. N., Puga, R., Stein, R. T., & Pitrez, P. M. (2016). Identifying a biomarker network for corticosteroid resistance in asthma from bronchoalveolar lavage samples. *Molecular Biology Reports*, *43*(7), 697–710. https://doi.org/10.1007/s11033-016-4007-x

Vilardell, M., Parra, G., & Civit, S. (2014). WISCOD: A Statistical Web-Enabled Tool for the Identification of Significant Protein Coding Regions. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/282343

Vlahou, A., Schorge, J. O., Gregory, B. W., & Coleman, R. L. (2003). Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data. *Journal of Biomedicine and Biotechnology*, *2003*(5), 308–314. https://doi.org/10.1155/S1110724303210032

Wan, F. C., Cui, Y. P., Wu, J. T., Wang, J. M., Liu, Q. Z., & Gao, Z. L. (2013). The PPI network and

cluster ONE analysis to explain the mechanism of bladder cancer. *European Review for Medical and Pharmacological Sciences*, *17*(5), 618–623.

Wang, J., Li, L., Yang, P., Chen, Y., Zhu, Y., Tong, M., Hao, Z., & Li, X. (2018). Identification of cervical cancer using laser-induced breakdown spectroscopy coupled with principal component analysis and support vector machine. *Lasers in Medical Science*, *33*(6), 1381–1386. https://doi.org/10.1007/s10103-018-2500-2

Wu, S. X., Wai, H. T., Li, L., & Scaglione, A. (2018). A Review of Distributed Algorithms for Principal Component Analysis. *Proceedings of the IEEE*, *106*(8), 1321–1340. https://doi.org/10.1109/JPROC.2018.2846568

Yin, S., Zheng, J., Jia, C., Zou, Q., Lin, Z., & Shi, H. (2022). UPFPSR: A ubiquitylation predictor for plant through combining sequence information and random forest. *Mathematical Biosciences and Engineering*, *19*(1), 775–791. https://doi.org/10.3934/mbe.2022035

Zhang, G., Si, Y., Wang, D., Yang, W., & Sun, Y. (2019). Automated Detection of Myocardial Infarction Using a Gramian Angular Field and Principal Component Analysis Network. *IEEE Access*, *7*, 171570–171583. https://doi.org/10.1109/ACCESS.2019.2955555

Zhang, Y. H., Hoopmann, M. R., Castaldi, P. J., Simonsen, K. A., Midha, M. K., Cho, M. H., Criner, G. J., Bueno, R., Liu, J., Moritz, R. L., & Silverman, E. K. (2021). Lung proteomic biomarkers associated with chronic obstructive pulmonary disease. *American Journal of Physiology - Lung Cellular and Molecular Physiology*, *321*(6), L1119–L1130. https://doi.org/10.1152/ajplung.00198.2021