JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org ISSN 2988-7003 (online)

Advances in Machine Learning and Deep Learning towards Medical Data Analysis

Andicha Vebiyatama[1] and Muji Ernawati[2]*

[1]*PT. Sysmex Indonesia, Indonesia*
[2] *PT. Ellison Global Indonesia, Indonesia*

ARTICLE HISTORY

Received: 26 March 24
Final Revision: 30 March 24
Accepted: 30 March 24
Online Publication: 31 March 24

KEYWORDS

Artificial Intelligence, Challenges, Deep Learning, Machine Learning, Medical Data Analysis

CORRESPONDING AUTHOR

andika.vebiyatama@gmail.com

DOI

10.37034/medinftech.v2i1.32

A B S T R A C T

Artificial intelligence uses advanced algorithms such as deep learning and machine learning methods to help doctors make more accurate diagnoses, identify potential health risks, and customize personalized treatment plans for patients. This literature review explores machine learning and deep learning methods applied to medical datasets over the past five years. The paper discusses the advancements, challenges, and future directions in utilizing ML and DL techniques for medical data analysis. It synthesizes recent research findings, highlighting key methodologies, datasets, and outcomes.

1. Introduction

Artificial intelligence (AI) is a term used to describe a machine or software's ability to mimic intelligent human behavior, perform quick calculations, solve problems, and analyze new data based on previously processed information [1]. AI significantly impacts various sectors, including agriculture, manufacturing, autonomous vehicles, fashion, sports analytics, healthcare, and the medical industry. In healthcare, AI has revolutionized the medical field, encompassing imaging and electronic medical records (EMR), lab diagnostics, treatments, boosting physician intelligence, new drug development, offering preventive and precision medicine, extensive biological data analysis, accelerating processes, and providing data storage and access for health organizations [1]. AI utilizes sophisticated algorithms and machine learning methods to assist doctors in making more precise diagnoses, spotting potential health risks, and tailoring personalized treatment plans for patients [2]. Deep learning has also been applied to segmentation and classification of organs and tumors of different types, and to predict treatment response or prognosis based on changes in tumor size or texture [3].

Machine learning, involving the utilization of algorithms to empower computers for more efficient task execution than humans by facilitating automatic data access, allows machines to accumulate experience and enhance performance over time, akin to deep learning, a predominant method in machine learning, which mirrors the cognitive development of infants, where billions of interconnected neurons adapt and modify pathways during information processing [4]. Machine learning and deep learning have been increasingly used in medical data analysis and healthcare applications. These techniques have the potential to improve the accuracy and efficiency of various diagnostic and treatment processes, as well as provide decision support to clinicians. Their applications in various healthcare scenarios, such as computer-aided medical diagnostics, drug discovery and development, medical imaging, automation, robotic surgery, electronic smart records creation, outbreak prediction, medical image analysis, and radiation treatments [5].

Numerous studies on medical data have been conducted using various methods such as the fuzzy c-means algorithm, K-means clustering, machine learning algorithms, deep learning algorithms, Genetic Algorithm (GA), and backpropagation network (BPN) [6]. For instance, machine learning algorithms like Support Vector Machine (SVM) [7], Random Forest [8], and K-Nearest Neighbour [9] have been employed. Additionally, the most popular deep learning algorithm in medical research is the Convolutional Neural Network (CNN) [10]. In recent times, the combination of machine learning and medicine has brought about

major improvements in diagnosing, predicting, and treating diseases. This paper seeks to offer a detailed review of the latest machine learning and deep learning methods used in analyzing medical data. The main objectives of this study include the datasets used, the commonly employed machine learning and deep learning techniques, and future research directions and challenges for future researchers.

This paper is organized as follows: Section II discusses the review of previous studies. Section III presents an overview of previous research and some directions for future research. Section V summarizes this paper.

2. Review of Previous Research

A systematic review approach was adopted to identify peer-reviewed papers published within the last five years. Criteria for inclusion/exclusion were established to ensure the selection of high-quality studies focusing on machine learning and deep learning methods applied to medical datasets. In the following discussion, several findings from the literature review regarding the utilization of machine learning and deep learning in the context of healthcare are presented.

This research yields significant findings in the development of hybrid feature selection algorithms for cancer diagnosis. This study aimed to evaluate Machine Learning models in predicting cancer with the highest accuracy using minimal features. By utilizing normalized microarray datasets, they performed data preprocessing and feature selection using various methods, including Chi-Squared, F-statistics, Mutual Information, as well as MI-based gene selection and Genetic Algorithm (GA). The research results indicated that the MI-GA gene selection method outperformed other methods, achieving maximum classification accuracy for various cancer datasets, such as leukemia, breast cancer, SRBCT, and lung cancer. However, the study also identified several shortcomings, particularly related to the specific characteristics of microarray data, which have high dimensions and limited sample sizes, making analysis challenging and presenting dimensionality issues in classification task implementation. Nevertheless, the method's strengths lie in its high performance and attractive time and cost efficiency, making it optimal for large-scale data [11].

This research discusses the use of machine learning algorithms in analyzing and detecting diabetes. They compare the performance of various models such as Random Forest, SVM, and KNN, emphasizing the importance of feature selection and cross-validation. In this study, data preprocessing is performed through normalization, outlier identification, and feature engineering, with the addition of exclusive features such as BMI and glucose categories. The results indicate that the use of optimal parameters, particularly in the Random Forest algorithm, improves model performance and reduces the risk of overfitting. However, these advantages are balanced by the complexity of the model, which may complicate interpretation and implementation. In the context of diabetes diagnosis, this research makes a significant contribution to the development of computer-aided systems for early diabetes detection, which can enhance the effectiveness of treatment and condition management [8].

The study aimed to perform a comparative study of various classification methods and feature selection techniques to predict diabetes with higher accuracy. They examined several classification algorithms, including multilayer perceptron, decision trees, K-nearest neighbor, and random forest classifiers, by applying various feature selection techniques. The research findings indicated that the best classification model was the random forest, which achieved an accuracy of 79.8%. The recommendation from this study is to use the random forest with six relevant features selected from correlation attribute evaluation for diabetes data classification. However, the study also identified some limitations, such as the use of mean imputation to fill in missing values, which may not always yield the best results, lack of discussion on the potential overfitting of the classification models used, and the limited generalization of results due to being conducted on only one dataset. Nonetheless, this research provides a significant contribution to the development of computer-aided systems for early detection of diabetes, which can enhance the effectiveness of treatment and condition management [9].

This study proposed a new architecture that combines K-means clustering and Support Vector Machine (SVM) techniques aimed to create a diabetes prediction model with the highest accuracy. This research achieved a very high accuracy of 98.7% on the Pima Indians Diabetes Database, indicating a significant improvement over previous methods. The results showed that the best classification model was random forest, but various other methods also achieved quite high accuracy, including some techniques that combine K-means clustering with other algorithms such as Genetic Algorithm (GA) and Correlation-based Feature Selection (CFS). However, the study also identified some weaknesses, including limitations in generalizing results due to focusing on one dataset and room for further improvement in the effectiveness of some methods and algorithms. Overall, this research makes a significant contribution to the development of diabetes prediction models that can enhance the management and treatment of the condition [7].

The study carried out aims to identify and analyze various methods and algorithms for breast tumor classification. This research involved the utilization of diverse machine learning techniques such as Support Vector Machine (SVM), K-Nearest Neighbors, and

image feature analysis techniques like Gray Level Co-Occurrence Matrix (GLCM). They utilized datasets from various sources, including the Wisconsin Breast Cancer dataset and datasets from leading medical institutions. The research findings indicated that SVM achieved the highest accuracy, reaching 97.13%, outperforming other classification models. The study also revealed that St-based linear SVM was suitable for small datasets, while SVM and C5.0 surpassed clustering models with an accuracy of 81%. Additionally, the research underscored the need for utilizing deep learning approaches to achieve higher accuracy, albeit requiring larger datasets and more resources. In conclusion, this study provides valuable insights into breast tumor classification methods, particularly highlighting the superiority of SVM, and emphasizes the importance of developing more advanced approaches to enhance the management and treatment of breast cancer [12].

The study aims to develop an effective cervical cancer prediction model with a focus on higher prediction accuracy. They proposed the Cervical Cancer Prediction Model (CCPM), which integrates outlier detection and data balancing techniques using density-based clustering with noise (DBSCAN), iForest, and synthetic minority oversampling techniques (SMOTE and SMOTETomek), along with random forest for cervical cancer prediction based on risk factors. The study utilized datasets from the Hospital Universitario de Caracas in Caracas, Venezuela, and the UCI repository, comprising 858 instances with 36 features relevant to cervical cancer. Various methods and algorithms, such as Support Vector Machine (SVM), K-Nearest Neighbors, Random Forest, Deep Learning (DL), and Genetic Algorithm (GA), were employed to process and analyze the datasets. The research findings indicated that the developed model achieved high prediction accuracy, with some technique combinations reaching accuracies of over 98%. Nevertheless, the study highlights the need for improvements in certain methods and algorithms, as well as the potential use of deep learning approaches to attain higher accuracy. In conclusion, this study provides a significant contribution to the development of cervical cancer prediction models that can enhance the management and treatment of this disease [13].

The research carried out aims to develop an accurate approach for diagnosing breast cancer by integrating machine learning techniques and feature selection/feature extraction. They proposed a hybrid approach that utilized Linear Discriminant Analysis (LDA) to reduce feature dimensions and then applied machine learning techniques. This research utilized a dataset from the Hospital Universitario de Caracas in Caracas, Venezuela, comprising 858 instances with 36 relevant features. Various methods and algorithms, such as Principal Component Analysis (PCA), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), among others, were employed to process and analyze the data. The results demonstrated that the proposed model could achieve a high level of accuracy in breast cancer diagnosis, surpassing previous studies. However, the study underscores the importance of continually improving the methods and algorithms used, as well as developing more diverse approaches for outlier detection and over-sampling methods. In conclusion, this research provides a significant contribution to the development of breast cancer diagnosis models that can enhance early diagnosis and management of the disease [14].

The study aims to develop artificial intelligence (AI)-based algorithms for the diagnosis and prognosis of kidney cancer. The research sought to compare the performance of AI-based predictors with non-AI ones, indicating a slightly better performance of AI predictors. However, the improvement in AI performance over the last decade has been relatively minor, and the number of genes used in the study had little influence on performance indices. The research methods included the use of various datasets, including methylation data, gene expression, and algorithms such as Convolutional Neural Networks (CNNs), Transformers, and Support Vector Machines (SVM). The results showed that AI models can predict kidney cancer stages well, but challenges remain, such as overfitting data, difficult-to-assess influences of each input variable, and the "black box" complexity of AI algorithms. Nevertheless, the use of DNA methylation data promises to improve predictor performance, and the study results indicated that the selected genes have a high correlation with overall survival, offering hope for further development in the diagnosis and prognosis of kidney cancer [15].

Table 1. Comparison of accuracy results from existing research

Ref.	Dataset	Feature Selection	Algorithm	Accuracy
[11]	Microarray Data	Gene selection method based on Mutual Information (MI)-Genetic Algorithm (GA)	SVM	100%
[8]	Diabetes Patient Data	Wrapper method	RFWBP	95,83%
[9]	Pima Indians Diabetes	Correlation Based Feature Selection, Principal Component Analysis	RF	79,8%
[7]	Pima Indians Diabetes	K-Means Clustering Algorithm	SVM	98,7%

Ref.	Dataset	Feature Selection	Algorithm	Accuracy
[12]	Wisconsin Diagnostic Breast Cancer (WDBC) dataset	Recursive feature elimination (RFE)	CNN	98,62%
[13]	Cervical Cancer Dataset	Chi-Squared Feature Selection	RF	98%
[14]	Wisconsin Breast Cancer Dataset	Correlation analysis and principal component analysis	Ensemble based learning	99%
[15]	TCGA data	Logistic regression	SVM	81.15%

3. Result and Discussion

Based on the review conducted, several key points will be discussed, including the types of data used, the machine learning and deep learning techniques applied, and potential future directions for further research.

3.1. Research datasets

A dataset is an accumulation of information extracted from previous data and is prepared for organization into fresh insights. Its purpose is to evaluate a research methodology crafted by experts, ensuring that the research is both comparable, repeatable, and verifiable [16]. Based on the results of the review of previous studies, there are various types of data and datasets used in those studies. Here are some key points that can be summarized from the research data:

Microarray Dataset: datasets consist of data collections related to cancer types such as ovarian cancer, small round blue cell tumor (SRBCT), lung cancer, and leukemia. This dataset undergoes normalization and processing using three filter methods: Chi-Square, F-statistics, and Mutual Information [11].

Diabetes Patient Data: This data involves diabetes patients aged 21-81 years. Preprocessing includes outlier identification, feature engineering, and feature selection. Exclusive features are added based on raw data features such as BMI category, glucose, blood, skin thickness, and insulin. One-hot encoding technique is also used to convert categorical variables into numerical form.

Various Dataset Sources: Datasets come from various sources including UCI Repository, health surveys, electronic health records, MRI scans, social media data, and specific medical institutions. This data is used to test and evaluate various models and algorithms in diagnosing and predicting diseases such as diabetes, breast cancer, cervical cancer, and others.

Use of DNA Methylation and Gene Expression Data: Some studies use DNA methylation and gene expression data to improve predictor performance in predicting diseases.

3.2. Machine Learning and Deep Learning Approaches in Medical Data Analysis

This section reviews various techniques used for medical data analysis based on observations from previous studies. The studies conducted used various machine learning and deep learning techniques and algorithms to analyze medical data and classify different health conditions, including cancer, diabetes, and obesity. The Table 1 highlights various algorithms used across different studies, each contributing to the accuracy and efficiency of medical data analysis. Support Vector Machine (SVM) is a supervised learning algorithm effective for classification and regression tasks, creating a hyperplane that separates data points into different classes with maximum margin. This algorithm was used effectively in the Microarray Data and Pima Indians Diabetes datasets. Random Forest (RF) is a popular ensemble learning method that combines multiple decision trees to improve model accuracy and robustness, used in the Microarray Data, Pima Indians Diabetes, Cervical Cancer, and Wisconsin Diagnostic Breast Cancer datasets, yielding high accuracy rates. RFWBP (Random Forest with Backpropagation), a combination of Random Forest and neural network techniques, was used in the Diabetes Patient Data study to leverage the strengths of both methods in enhancing predictive performance. Convolutional Neural Network (CNN) is a deep learning algorithm particularly effective for image and pattern recognition tasks, used in the Wisconsin Diagnostic Breast Cancer dataset and demonstrating high performance and accuracy. Ensemble Based Learning combines multiple models to improve predictive accuracy and generalization, applied in the Wisconsin Breast Cancer dataset and resulting in high accuracy rates. K-Means Clustering Algorithm was used in the Pima Indians Diabetes study, grouping data points into clusters based on their similarities, assisting in feature selection and data segmentation. These algorithms are selected based on the specific needs and characteristics of each dataset, aiming to optimize model performance and predictive accuracy. The combination of various methods across different studies demonstrates the versatility and effectiveness of modern machine learning and data analysis techniques.

Feature selection is a critical process in machine learning and data analysis that helps identify and select the most relevant features from a dataset, thereby enhancing model performance and efficiency [17]. In the studies presented in the Table 1, various feature selection methods were utilized: the gene selection method based on Mutual Information (MI)-Genetic Algorithm (GA) in the Microarray Data dataset combines mutual information and genetic algorithms to optimize model performance [11]; the wrapper method in the Diabetes Patient Data study evaluates different subsets of features and selects the ones that yield the

best model performance [8]; correlation-based feature selection and principal component analysis were used in the Pima Indians Diabetes dataset to identify the most relevant features based on their correlation and data variance [9]; recursive feature elimination (RFE) in the Wisconsin Diagnostic Breast Cancer dataset recursively removes the least important features to enhance overall model accuracy [12]; Chi-Squared feature selection in the Cervical Cancer dataset selects features based on their association with the target variable [13]; correlation analysis and principal component analysis in the Wisconsin Breast Cancer dataset identify the most significant features based on their relationships with the target variable and data variance [14]; and logistic regression, primarily a modeling technique, was used in the TCGA data study for feature selection by identifying important variables contributing significantly to the model [15]. These feature selection methods are crucial in improving model performance by focusing on the most significant features, reducing dimensionality, and potentially lowering computational costs.

The best results presented in the Table 1 show several studies with very high accuracy rates. In the Microarray Data dataset, the feature selection method based on Mutual Information (MI) and Genetic Algorithm (GA) with SVM algorithms achieved an accuracy rate of 100% [11]. Additionally, in the Wisconsin Breast Cancer dataset, correlation analysis and principal component analysis with ensemble-based learning achieved an accuracy rate of 99% [14]. These studies demonstrate the effectiveness of the methods and algorithms used in predicting medical outcomes very well, especially in studies involving advanced feature selection such as MI-GA and other optimization techniques.

3.3. Challenges and Future Directions

Despite the promising results achieved by machine learning and deep learning techniques in medical data analysis, several challenges persist. Issues related to data quality, interpretability, and model generalizability are discussed in this section. Additionally, ethical considerations, such as patient privacy and algorithmic bias, are addressed, emphasizing the importance of responsible AI deployment in healthcare settings. These include the development of interpretable ML models, integration of heterogeneous data sources, and exploration of federated learning approaches to facilitate collaborative research while ensuring data privacy. To overcome current challenges and further enhance the utility of machine learning and deep learning in medical research, this section proposes several future directions.

1. Limitations on the Used Dataset: Most studies utilize specific datasets, such as the Pima Indians Diabetes dataset. This may result in an inability to generalize the findings to broader populations. The solution is to use more diverse and population-representative datasets or validate the results on different datasets to confirm findings.

2. Potential Overfitting: Some studies do not discuss in detail the potential overfitting of the models used. Overfitting can result in models that cannot generalize well to new data. The solution is to use regularization techniques, such as dropout or L1/L2 penalties, and perform cross-validation to objectively evaluate model performance like K-Fold validation [18].

3. Difficulty in Model Interpretation: Some studies face complexity in interpreting models, especially when using more complex models like Random Forest. The solution is to use model interpretation techniques such as feature importance or SHAP values, and simplify the model if possible to improve interpretability.

4. Limitations on the Used Data Imputation Methods: Some studies use simple data imputation methods like mean imputation, which may not always provide the best results and can affect data distribution. The solution is to use more advanced imputation methods, such as imputation using machine learning models or imputation techniques that consider the existing data structure.

5. Limitations on Model Evaluation: Some studies may not comprehensively evaluate the models, especially in terms of generalization of results and evaluating model performance on unseen data. The solution is to use comprehensive model evaluation techniques, such as appropriate metrics usage, cross-validation, and testing on external datasets.

By addressing these shortcomings, research in the field of machine learning in healthcare can become more reliable and dependable in aiding diagnosis and disease prediction effectively.

4. Conclusion

This paper provides a comprehensive overview of recent advancements in machine learning and deep learning methods applied to medical datasets. It underscores the potential of these techniques to revolutionize healthcare delivery, improve diagnostic accuracy, and enhance patient outcomes. Overall, these studies demonstrate the use of various types of data and preprocessing techniques implemented to produce models and algorithms effective in diagnosing and predicting various health conditions, including cancer, diabetes, and obesity. SVM with optimal parameters achieved maximum accuracy in detecting diabetes patients and classifying breast cancer and SRBCT datasets. By addressing existing challenges and embracing future opportunities, the integration of AI into medical practice holds promise for transformative impact in the years to come.

References

[1] D. D. Farhud and S. Zokaei, "Ethical issues of artificial intelligence in medicine and healthcare," *Iran. J. Public Health*, vol. 50, no. 11, pp. i–v, 2021, doi: 10.18502/ijph.v50i11.7600.

[2] M. R. King, "The Future of AI in Medicine: A Perspective from a Chatbot," *Ann. Biomed. Eng.*, vol. 51, no. 2, pp. 291–295, 2023, doi: 10.1007/s10439-022-03121-w.

[3] H. P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep Learning in Medical Image Analysis," *Adv. Exp. Med. Biol.*, vol. 1213, no. 3–21, pp. 1–26, 2020, doi: 10.1007/978-3-030-33128-3_1.

[4] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5: 91, pp. 1–16, 2023, doi: 10.3390/computers12050091.

[5] O. P. Jena, B. Bhushan, and U. Kose, *Machine Learning and Deep Learning in Medical Data Analytics and Healthcare Applications*, 1st ed. Boca Raton: CRC Press, 2022. doi: https://doi.org/10.1201/9781003226147.

[6] R. An, J. Shen, and Y. Xiao, "Applications of Artificial Intelligence to Obesity Research: Scoping Review of Methodologies," *J. Med. Internet Res.*, vol. 24, no. 12, pp. 1–32, 2022, doi: 10.2196/40589.

[7] N. Arora, A. Singh, M. Z. N. Al-Dabagh, and S. K. Maitra, "A Novel Architecture for Diabetes Patients' Prediction Using K - Means Clustering and SVM," *Math. Probl. Eng.*, vol. 2022, pp. 1–9, 2022, doi: 10.1155/2022/4815521.

[8] M. S. Ali, M. K. Islam, A. A. Das, D. U. S. Duranta, M. F. Haque, and M. H. Rahman, "A Novel Approach for Best Parameters Selection and Feature Engineering to Analyze and Detect Diabetes: Machine Learning Insights," *Biomed Res. Int.*, vol. 2023, pp. 1–15, 2023, doi: 10.1155/2023/8583210.

[9] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/3820360.

[10] H. Cho *et al.*, "Machine Learning and Health Science Research: Tutorial," *J. Med. Internet Res.*, vol. 26, pp. 1–15, 2024, doi: 10.2196/50890.

[11] T. Elemam and M. Elshrkawey, "A Highly Discriminative Hybrid Feature Selection Algorithm for Cancer Diagnosis," *Sci. World J.*, vol. 2022, pp. 1–15, 2022, doi: 10.1155/2022/1056490.

[12] S. H. Abdulla, A. M. Sagheer, and H. Veisi, "Breast Cancer Classification Using Machine Learning Algorithms: A Review," *Turkish J. Comput. Math. Educ.*, vol. 141, no. 14, pp. 1970–1979, 2021, doi: 10.1007/978-981-15-7106-0_56.

[13] M. F. Ijaz, M. Attique, and Y. Son, "Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods," *Sensors*, vol. 20, no. 10, pp. 1–23, 2020, doi: https://doi.org/10.3390/s20102809.

[14] S. Ibrahim, S. Nazir, and S. A. Velastin, "Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis," *J. Imaging*, vol. 7, no. 225, pp. 1–16, 2021, doi: 10.3390/jimaging7110225.

[15] M. Giulietti *et al.*, "The role of artificial intelligence in the diagnosis and prognosis of renal cell tumors," *Diagnostics*, vol. 11, no. 206, pp. 1–12, 2021, doi: 10.3390/diagnostics11020206.

[16] M. Ernawati, E. H. Hermaliani, and D. N. Sulistyowati, "Penerapan DeLone and McLean Model untuk Mengukur Kesuksesan Aplikasi Akademik Mahasiswa Berbasis Mobile," *J. IKRA-ITH Inform.*, vol. 5, no. 18, pp. 58–67, 2020.

[17] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 52, pp. 1–26, 2020, doi: 10.1186/s40537-020-00327-4.

[18] S. Nuarini and A. Rumintarsih, "Optimization of Breast Cancer Prediction using Optimaze Parameter on Machine Learning," *J. Med. Informatics Technol.*, vol. 1, no. 1, pp. 25–30, 2023, doi: 10.37034/medinftech.v1i1.5.