Disease Prediction Using Machine Learning

Preprint · December 2020

DOI: 10.13140/RG.2.2.18279.47521

1 author:

Marouane Ferjani
Bournemouth University
2 PUBLICATIONS **3** CITATIONS

SEE PROFILE

Disease Prediction Using Machine Learning

*Research Gate Link:

Marouane Fethi Ferjani
Computing Department
Bournemouth University
Bournemouth, England
s5319941@bournemouth.ac.uk

***Abstract*—The wide adaptation of computer-based technology in the health care industry resulted in the accumulation of electronic data. Due to the substantial amounts of data, medical doctors are facing challenges to analyze symptoms accurately and identify diseases at an early stage. However, supervised machine learning (ML) algorithms have showcased significant potential in surpassing standard systems for disease diagnosis and aiding medical experts in the early detection of high-risk diseases. In this literature, the aim is to recognize trends across various types of supervised ML models in disease detection through the examination of performance metrics. The most prominently discussed supervised ML algorithms were Naïve Bayes (NB), Decision Trees (DT), K-Nearest Neighbor (KNN). As per findings, Support Vector Machine (SVM) is the most adequate at detecting kidney diseases and Parkinson's disease. The Logistic Regression (LR) performed highly at the prediction of heart diseases. Finally, Random Forest (RF), and Convolutional Neural Networks (CNN) predicted in precision breast diseases and common diseases, respectively.**

***Keywords*—Health Care, Supervised Machine Learning, Diseases Prediction**

I. Introduction

A. Motivation

The emergence of Artificial Intelligence (AI) enabled computerized systems to perceive, think and operate in an intelligent manner like humans [1]. AI is a multidisciplinary concept of ML, Computer Vision, Deep Learning, and Natural Language Processing [2]. ML algorithms apply various optimization, statistical ,and probabilistic techniques to learn from data that was generated from past experiences, and deploy it in decision making [3]. These algorithms deemed to be applied in many disciplines including network intrusion recognition, customer purchase behavior detection, process manufacturing optimization, credit card fraud detection, and disease modulation. Many of these applications have been designed using the supervised learning approach. In this approach, datasets with known labels are induced to prediction models to predict unlabeled examples [2], [3]. This presents the hypothesis that medical doctors can utilize supervised learning as a powerful tool to conduct diseases diagnosis more efficiently [4] .

Medicaid services and centers for Medicare reported that 50% of Americans had multiple chronic diseases, which led the US health care to spend around $3.3 trillion in 2016, that amounts to $10,348 per person in the US [5]. Moreover, the World Health Organization and World Economic Forum reported that India had a huge loss of $236.6 billion by 2015 because of fatal diseases, caused by malnutrition and morbid lifestyles [6]. Such expenditures revealed how prone people are to a spectrum of diseases, which showcased how vital it is to detect diseases early, to consequently reduce the fatality of these maladies. In addition, early disease prediction can lessen the financial pressure on the economy and ensure better maintenance on the overall well-being of the community [5], [6].

According to Yuan [7], ML algorithms are highly susceptible to errors because of two factors. Firstly, it depends on the quality and the selection of the datasets, which is crucial to achieve accurate and unbiased decisions. Secondly, ML algorithms relies heavily on the right selection of features extracted from the dataset, which proved to be difficult, time consuming, and required high computational power. These factors hinder the performance of the learning model and generate fatal errors that can endanger the lives of patients. In contrast, Ismaeel [8] argued that standard statistical techniques, the work experience and the intuition of medical doctors led to undesirable biases and errors when detecting risks associated to the disease. With the substantial surge of electronic health data, medical doctors are facing challenges to identify diseases accurately at an early stage. For this reason, advanced computational methodologies such as ML algorithms were introduced to discover meaningful patterns and hidden information from data, which can be used for critical decision making. In consequence, the burden on the medical staff decreased, while the survival rate of patients was ameliorated [3], [8].

B. Aim

The aim of this study is to test the proposed hypothesis that supervised ML algorithms can improve health care by the accurate and early detection of diseases. In this study, we investigate studies that utilize more than one supervised ML model for each disease recognition problem. This approach renders more comprehensiveness and precision because the evaluation of the performance of a single algorithm over various study settings induces bias which generates imprecise results. The analysis of ML models will be conducted on few diseases located at heart, kidney, breast, and brain. For the detection of the disease, numerous methodologies will be evaluated such as KNN, NB, DT, CNN, SVM, and LR. At

the end of this literature, the best performing ML models in respect of each disease will be concluded.

II. Literature Review

A. Common Diseases

Dahiwade et al. [9] proposed a ML based system that predicts common diseases. The symptoms dataset was imported from the UCI ML depository, where it contained symptoms of many common diseases. The system used CNN and KNN as classification techniques to achieve multiple diseases prediction. Moreover, the proposed solution was supplemented with more information that concerned the living habits of the tested patient, which proved to be helpful in understanding the level of risk attached to the predicted disease. Dahiwade et al. [9] compared the results between KNN and CNN algorithm in terms of processing time and accuracy. The accuracy and processing time of CNN were 84.5% and 11.1 seconds, respectively. The statistics proved that KNN algorithm is under performing compared to CNN algorithm. In light of this study, the findings of Chen et al. [10] also agreed that CNN outperformed typical supervised algorithms such as KNN, NB, and DT. The authors concluded that the proposed model scored higher in terms of accuracy, which is explained by the capability of the model to detect complex nonlinear relationships in the feature space. Moreover, CNN detects features with high importance that renders better description of the disease, which enables it to accurately predict diseases with high complexity [9], [10]. This conclusion is well supported and backed with empirical observations and statistical arguments. Nonetheless, the presented models lacked details, for instance, Neural Networks parameters such as network size, architecture type, learning rate and back propagation algorithm, etc. In addition, the analysis of the performances is only evaluated in terms of accuracy, which debunks the validity of the presented findings [9]. Moreover, the authors did not take into consideration the bias problem that is faced by the tested algorithms [9], [10]. In illustration, the incorporation of more feature variables could immensely ameliorate the performance metrics of under performed algorithms [11].

B. Kidney Diseases

Serek et al. [12] planned a comparative study of classifiers performance for Chronic Kidney disease (CKD) detection using The Kidney Function Test (KFT) dataset. In this study, the classifiers used are KNN, NB, and RF classifier; their performance is examined in terms of F-measure, precision, and accuracy. As per analysis, RF scored better in phrases of F-measure and accuracy, while NB yielded better precision. In consideration of this study, Vijayarani [13] aimed to detect kidney diseases using SVM and NB. The classifiers were used to identify four types of kidney diseases namely Acute Nephritic Syndrome, Acute Renal Failure, Chronic Glomerulonephritis, and CKD. Additionally, the research was focused on determining the better performing classification algorithm based on the accuracy and execution time. From the results, SVM considerably achieved higher accuracy than NB, which makes it the better performing algorithm. However, NB classified data with minimum execution time. Other several empirical studies also focused on locating CKD; Charleonnan et al. [14] and Kotturu et al. [15] concluded that the SVM classifier is the most adequate for kidney diseases because it deals well with semi-structured and unstructured data. Such flexibility allowed SVM to handle larger features spaces, which resulted in acquiring high accuracy when detecting complex kidney diseases. Although supported by findings, the conclusion is weakened by prior suggestion that different hyper-parameters were not experimented when evaluating the performances of ML algorithms. According to Uddin [3] the exploration of the hyper-parameter space can generate different accuracy results and render better performances for ML algorithms.

C. Heart Diseases

Marimuthu et al. [16] aimed to predict heart diseases using supervised ML techniques. The authors structured the attributes of data as gender, age, chest pain, gender, target and slope [16]. The applied ML algorithms that were deployed are DT, KNN, LR and NB. As per analysis, the LR algorithm gave a high accuracy of 86.89%, which deemed to be the most effective compared to the other mentioned algorithms. In 2018, Dwivedi [17] attempted to add more precision to the prediction of heart diseases by accounting for additional parameters such as Resting blood pressure, Serum Cholesterol in mg/dl, and Maximum Heart Rate achieved. The used dataset was imported from the UCI ML laboratory; it was comprised with 120 samples that were heart disease positive, and 150 samples that were heart disease negative. Dwivedi attempted to evaluate the performance of Artificial Neural Networks (ANN), SVM, KNN, NB, LR and Classification Tree. At the appliance of tenfold cross validation, the results showed that LR has the highest classification accuracy and sensitivity, which shows high dependability at detecting heart diseases [17]. This conclusion is strengthened by the findings of Polaraju [18] and Vahid et al. [19], where the Logistic Regression outperformed other techniques such as ANN, SVM, and Adaboost. The studies excelled in conducting an extensive analysis on the ML models. For instance, various hyper-parameters were tested at each ML algorithm to converge to the best possible accuracy and precision values. Despite that advantage, the small size of the imported datasets constraints the learning models from targeting diseases with higher accuracy and precision.

D. Breast Diseases

Shubair [20] attempted for the detection of breast cancer using ML algorithms, namely RF, Bayesian Networks and SVM. The researchers obtained the Wisconsin original breast cancer dataset from the UCI Repository and utilized it for comparing the learning models in terms of key parameters such as accuracy, recall, precision, and area of ROC graph. The classifiers were tested using K-fold validation method, where the chosen value of K is equal to 10 [20]. The simulation results have proved that SVM excelled in terms of recall, accuracy, and precision. However, RF had a higher probability

in the correct classification of the tumor, which was implied by the ROC graph. In contrast, Yao [21] experimented with various data mining methods including RF and SVM to determine the best suited algorithm for breast cancer prediction. Per results, the classification rate, sensitivity, and specificity of Random Forest algorithm were 96.27%, 96.78%, and 94.57%, respectively, while SVM scored an accuracy value of 95.85%, a sensitivity of 95.95%, and a specificity of 95.53%. Yao came to the conclusion that the RF algorithm performed better than SVM because the former provides better estimates of information gained in each feature attribute. Furthermore, RF is the most adequate at breast diseases classification, since it scales well for large datasets and prefaces lower chances of variance and data overfitting [21]. the studies advantageously presented multiple performance metrics that solidified the underlined argument. Nevertheless, the inclusion of the preprocessing stage to prepare raw data for training proved to be disadvantageous for ML models [21]. According to Yao [21], omitting parts of data reduces the quality of images, and therefore the performance of the ML algorithm is hindered.

E. Parkinson's Disease

Chen et al. [22] presented an effective diagnosis system using Fuzzy k-Nearest Neighbor (FKNN) for the diagnosis of Parkinson's disease (PD) . The study focused on comparing the proposed SVM-based and the FKNN-based approaches. the Principal Component Analysis (PCA) was utilized to assemble the most discriminated features for the construction of an optimal FKNN model. The dataset was taken from the UCI depository, and it recorded numerous biomedical voice measurement ranging from 31 people, 24 with PD. The experimental findings have indicated that the FKNN approach advantageously achieves over the SVM methodology in terms of sensitivity, accuracy, and specificity. In line of this study, Behroozi [23] aimed to propose a new classification framework to diagnose PD, which was enhanced by a filter-based feature selection algorithm that increased the classification accuracy up to 15%. The classification of the framework was characterized by applying independent classifiers for each subset of the dataset to account for the loss of valuable information. The chosen classifiers were KNN, SVM, Discriminant Analysis and NB. The results showed that SVM achieved the highest in all the performance metrics. In addition, Eskidere [24] concentrated on tracking the progression of PD by discussing the performance of SVM with other classifiers such as Least Square Support Vector (LS-SVM), General Regression Neural Network (GRNN) and Multi-layer Perceptron Neural Network (MLPNN). The findings indicated that LS-SVM is the highest performing model. This conclusion is strengthened by the adequate comparison of decoders with their optimal performance metric [25]. According to Lavesson [25], various ML algorithms are designed to optimize numerous performance metrics (e.g., Neural Networks optimizes squared error whereas KNN and SVM optimize accuracy). Furthermore, the authors are particularly good at proposing frameworks with details. For example, SVMs parameters such as the kernel and the regularization value were outlined in depth. However, ML models were not calibrated before evaluating the performances. Caruana argues that [26] calibration substantially enhances the classification of few learning models namely NB, SVM, and RF.

CONCLUSION

The use of different ML algorithms enabled the early detection of many maladies such as heart, kidney, breast, and brain diseases. Throughout the literature, SVM, RF and LR algorithms were the most widely used at prediction, while accuracy was the most used performance metric. The CNN model proved to be the most adequate at predicting common diseases. Furthermore, SVM model showed superiority in accuracy at most times for kidney diseases and PD because of its reliability in handling high-dimensional, semi-structured and unstructured data. For Breast cancer prediction, RF showed more superiority in the probability of correct classification of the diseases because of its ability to scale well for large datasets and its susceptibility to avoid overfitting. Finally, the LR algorithm proved to be the most reliable in predicting heart diseases.

In future work, the creation of more complex ML algorithms is much needed to increase the efficiency of disease prediction. In addition, learning models should be calibrated more often after the training phase for potentially a better performance. Moreover, datasets should be expanded on different demographics to avoid overfitting and increase the accuracy of the deployed models. Finally, more relevant feature selection methods should be used to enhance the performance of the learning models.

REFERENCES

[1] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 1275–1278.

[2] Y. Hasija, N. Garg, and S. Sourav, "Automated detection of dermatological disorders through image-processing and machine learning," in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 2017, pp. 1047–1051.

[3] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–16, 2019.

[4] R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: A survey," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 302–305.

[5] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 2018, pp. 1–4.

[6] M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra, "A proposed model for lifestyle disease prediction using support vector machine," in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2018, pp. 1–6.

[7] F. Q. Yuan, "Critical issues of applying machine learning to condition monitoring for failure diagnosis," in *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2016, pp. 1903–1907.

[8] S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning machine (elm) technique for heart disease diagnosis," in *2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015)*, 2015, pp. 1–3.

[9] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019*, no. Iccmc, pp. 1211–1215, 2019.

[10] S. Jadhav, R. Kasar, N. Lade, M. Patil, and S. Kolte, "Disease Prediction by Machine Learning from Healthcare Communities," *International Journal of Scientific Research in Science and Technology*, pp. 29–35, 2019.

[11] R. Saravanan and P. Sujatha, "A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 945–949.

[12] Y. Amirgaliyev, S. Shamiluulu, and A. Serek, "Analysis of chronic kidney disease dataset by applying machine learning methods," in *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, 2018, pp. 1–4.

[13] V. S and D. S, "Data Mining Classification Algorithms for Kidney Disease Prediction," *International Journal on Cybernetics & Informatics*, vol. 4, no. 4, pp. 13–25, 2015.

[14] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," *2016 Management and Innovation Technology International Conference, MITiCON 2016*, pp. MIT80–MIT83, 2017.

[15] P. Kotturu, V. V. Sasank, G. Supriya, C. S. Manoj, and M. V. Maheshwarredy, "Prediction of chronic kidney disease using machine learning techniques," *International Journal of Advanced Science and Technology*, vol. 28, no. 16, pp. 1436–1443, 2019.

[16] M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach," *International Journal of Computer Applications*, vol. 181, no. 18, pp. 20–25, 2018.

[17] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, no. 10, pp. 685–693, 2018.

[18] K. Polaraju, D. Durga Prasad, and M. Tech Scholar, "Prediction of Heart Disease using Multiple Linear Regression Model," *International Journal of Engineering Development and Research*, vol. 5, no. 4, pp. 2321–9939, 2017. [Online]. Available: www.ijedr.org

[19] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *2017 IEEE Symposium on Computers and Communications (ISCC)*, 2017, pp. 204–207.

[20] P. P. Sengar, M. J. Gaikwad, and A. S. Nagdive, "Comparative study of machine learning algorithms for breast cancer prediction," *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, pp. 796–801, 2020.

[21] D. Yao, J. Yang, and X. Zhan, "A novel method for disease prediction: Hybrid of random forest and multivariate adaptive regression splines," *Journal of Computers (Finland)*, vol. 8, no. 1, pp. 170–177, 2013.

[22] H. L. Chen, C. C. Huang, X. G. Yu, X. Xu, X. Sun, G. Wang, and S. J. Wang, "An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach," *Expert Systems with Applications*, vol. 40, no. 1, pp. 263–271, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2012.07.014

[23] M. Behroozi and A. Sami, "A multiple-classifier framework for Parkinson's disease detection based on various vocal tests," *International Journal of Telemedicine and Applications*, vol. 2016, 2016.

[24] Ö. Eskidere, F. Ertaş, and C. Hanilçi, "A comparison of regression methods for remote tracking of Parkinson's disease progression," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5523–5528, 2012.

[25] N. Lavesson, *Evaluation and Analysis of Supervised Learning Algorithms and Classifiers*, 2006.

[26] R. Caruana and A. Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics," *Proceedings of the 23rd international conference on Machine Learning*, pp. 161–168, 2006. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.3232