OPTIMIZING DIABETES DIAGNOSIS: SYSTEMATIC REVIEW OF FEATURE SELECTION FOR PREDICTIVE MODELING

by

ANGELA C. MUNOZ

B.A., Georgia College & State University, 2008

M.S.I.T., Middle Georgia State University, 2021

A Research Paper Submitted to the School of Computing Faculty of

Middle Georgia State University in

Partial Fulfillment for the Requirements for the Degree

DOCTOR OF SCIENCE IN INFORMATION TECHNOLOGY

MACON, GEORGIA

2024

Optimizing diabetes diagnosis: systematic review of feature selection for predictive modeling

Angela Munoz, *Middle Georgia State University, angela.munoz@mga.edu*

Abstract

The escalating global prevalence of diabetes necessitates transformative advancements in diagnostic methodologies. This systematic review evaluates feature selection (FS) techniques for predictive modeling, emphasizing their crucial role in enhancing accuracy and efficiency. Synthesizing literature on machine learning in healthcare, the study underscores FS's foundational importance in refining predictive models for diabetes diagnosis. Key findings highlight the necessity of tailored FS methodologies and the integration of machine learning algorithms to optimize predictive modeling accuracy. This review offers insights into the current landscape of FS techniques and provides valuable directions for future research, contributing to the advancement of precise and efficient predictive models in diabetes diagnosis, crucial in the context of machine learning applications in healthcare.

Keywords: feature selection, predictive modeling, machine learning, precision medicine, global health

Introduction

Projections suggest by 2045, the number of individuals affected by diabetes is poised to surge by 46% to reach an alarming 783 million people worldwide (World Health Organization, n.d.). Despite the potential of disease-specific datasets and advanced computational methodologies, a transformative opportunity exists for the diagnosis and treatment of diabetes. This knowledge empowers healthcare practitioners to interpret results from diagnostic tests, making informed decisions and elevating the quality of care (Shreffler, 2023). This systematic review addresses a critical gap by comprehensively assessing the impact of "feature selection" techniques on predictive modeling for diabetes diagnosis, aiming to inform healthcare practitioners, researchers, and data scientists. The study contributes to enhancing the quality of care for individuals affected by diabetes, aligning with the World Health Organization's goal to combat the rising global burden of this disease.

Problem statement

This study addresses the gap in understanding the influence of "feature selection" techniques in the literature on diabetes diagnosis and predictive modeling. While feature selection (FS) is crucial for enhancing predictive model accuracy and efficiency, its untapped potential for improving diabetes diagnosis needs exploration. Shilaskar and Ghatol (2013) emphasize the importance of diverse features for complex medical problem diagnosis, and Ershadi and Seifi (2022) stress feature reduction and selection's role in enhancing classifier performance. Alhassan and Wan Zainon (2021) call for research focusing on new hybrid classification techniques to improve accuracy and computational effectiveness in chronic disease detection. Addressing this research gap is crucial for advancing our understanding of feature selection's impact on diabetes diagnosis and predictive modeling.

Purpose of the study

The purpose of this study is to systematically review and evaluate the current state of research on FS for predictive modeling in diabetes diagnosis. By synthesizing available evidence, the study aims to identify the most effective FS techniques and their impact on the accuracy and reliability of predictive models in diabetes diagnosis.

Research question

RQ1: What are the predominant themes within the literature on FS techniques, and how do these themes influence the accuracy and reliability of predictive models in diabetes diagnosis?

Research objectives

This research project has two significant objectives. First, the paper reviews and synthesizes FS articles in predictive modeling for diabetes diagnosis, highlighting the most popular approaches. Secondly, the study aims to improve the prediction of diabetes diagnosis.

Review of the literature

Importance of feature selection

The critical significance of feature selection (FS) is underscored by Balogun et al. (2020), who emphasized its paramount role in addressing high dimensionality issues within software defect prediction. Through an extensive benchmark study involving 46 FS methods and 25 defect datasets sourced from major repositories, their objective was to bring clarity and resolution to contradictions present in existing studies. Additionally, Battineni et al. (2020) highlighted the substantial importance of selecting pertinent features in machine learning, specifically for chronic illness diagnosis and healthcare decision-making. They accentuated the pivotal role of feature selection in achieving precision within predictive modeling. Furthermore, Kanyongo and Ezugwu (2023) shed light on the necessity of extending feature selection methodologies to diverse healthcare contexts, placing particular emphasis on the imperative for additional research concerning African electronic health record databases. Balogun et al.'s (2020) extensive empirical study, focused on software defect prediction and feature selection, serves as a cornerstone, emphasizing the profound impact the choice of feature selection methods, classifiers, and datasets can have on model performance. The collective recognition of FS as pivotal in enhancing predictive models resonates across these studies, transcending various domains.

Machine learning in healthcare

In the realm of predicting medication adherence for non-communicable diseases, Kanyongo and Ezugwu (2023) showcased the indispensable role played by machine learning. Spencer et al. (2020) accentuated the advantages of amalgamating feature selection with machine learning algorithms to predict heart disease. Additionally, Chaudhuri et al. (2021) introduced a multi-stage approach aimed at elevating the reliability and accuracy of cervical cancer diagnosis through the application of machine learning techniques. Collectively, these studies emphasize and illuminate the transformative impact machine learning imparts within healthcare domains.

Hybrid approaches

Jain and Singh (2018) systematically categorized feature selection algorithms and advocated for research into hybrid classification approaches, aiming to enhance both classifier accuracy and computational efficiency. Their efforts were aligned with broader objectives focused on improving predictive models, with a specific emphasis on chronic disease prediction. In a related study, Rauber et al. (2015) presented experimental findings using Case Western Reserve University Bearing Data. They demonstrated the effectiveness of their strategy, which involved integrating various feature models within a unified pool. Through a combination of feature selection, this approach optimized the fault diagnosis system, effectively filtering out discriminative features. Notably, Rauber et al. introduced robust performance estimation techniques not commonly encountered in the field of engineering. The implementation of hybrid approaches, in line with Jain and Singh's recommendations (2018) and exemplified by Rauber et al. (2015), exhibits significant promise in improving classifier accuracy.

Metrics for evaluation

The evaluation of healthcare predictive models necessitates the use of appropriate metrics, as emphasized by Mohammad et al. (2022) in their focus on software fault prediction. They highlighted the crucial role of performance indicators such as accuracy, recall, and F1-measure. Similarly, Olivera et al. (2017) advocated for the utilization of the area under the curve (AUC) to assess the predictive effectiveness in diagnosing undiagnosed diabetes. Hassan et al. (2020) reiterated the importance of employing performance metrics in healthcare, presenting case studies utilizing machine learning for diagnosis and therapy, leveraging extensive healthcare datasets. Noteworthy is the study by Wu et al. (2010) which successfully predicted heart failure over six months before clinical diagnosis, underscoring the significance of rigorous evaluation through relevant metrics. The incorporation of diverse metrics, as recommended by Mohammad et al. (2022) and demonstrated by Wu et al. (2010), enhances the robustness of evaluation in healthcare predictive modeling.

Challenges and opportunities in healthcare data

Patiño-Saucedo et al. (2022) delved into the challenges associated with addressing domain-specific feature selection and optimization difficulties crucial for accurate healthcare diagnosis. In a complementary study, Reps et al. (2018) introduced a structured and standardized framework for developing patient-level prediction models utilizing observational healthcare data. Their work addressed the pressing need for a transparent and reproducible approach in healthcare data analysis. Furthermore, Bashir et al. (2019) confronted the challenge of precisely predicting heart disease diagnoses in the medical field, emphasizing the pivotal role of data science as a crucial tool for early prediction and addressing large-scale data problems within the healthcare domain. Collectively, these studies illuminate the intricate challenges and significant opportunities in healthcare data analysis, underscoring the importance of adopting transparent and reproducible approaches.

Disease diagnosis and prediction

The pivotal role of features in disease prediction is evident in the noteworthy contributions of Kanyongo and Ezugwu (2023), who focused on noncommunicable diseases, and Spencer et al. (2020), who specialized in heart disease prediction. Rauber et al. (2015) significantly enhanced diagnostic accuracy by employing multiple feature extraction methods and feature selection, introducing an innovative strategy applicable to various disease diagnoses. Reddy et al. (2019) addressed the global issue of heart disease through the utilization of machine learning techniques, emphasizing the importance of feature selection methods for cost-effective diagnosis and the identification of key attributes crucial for predicting heart disease. Sakri et al. (2018) introduced particle swarm optimization as a feature selection technique, seamlessly integrated into three well-known classifiers—naive Bayes, K-nearest neighbor, and fast decision tree learner. Their primary goal was to augment the accuracy of prediction models, providing patients with more reliable and timely predictions. Stiglic et al. (2020) made valuable contributions by categorizing interpretability approaches into two main groups: personalized interpretation (local interpretability) and summarization of prediction models on a population level (global interpretability). Additionally, they discerned model-specific techniques tailored for interpreting predictions from specific models, such as neural networks, and model-agnostic approaches provide explanations for predictions made by any machine learning model. In a study by Wu et al. (2010), the authors successfully predicted heart failure over six months before clinical diagnosis, achieving an AUC of about 0.76 using logistic regression and Boosting, even with stringent model selection criteria. Conversely, SVM exhibited poorer performance, potentially attributable to imbalanced data. Collectively, these studies contribute to a nuanced understanding of disease diagnosis and prediction, emphasizing diverse approaches and considerations.

Variable selection methods for clinical prediction models

Bagherzadeh-Khiabani et al. (2016) underscored the pivotal role of employing variable selection methods derived from data mining to enhance the accuracy and reliability of clinical prediction models for diabetes diagnosis. Their insights strongly advocate for the necessity of optimizing predictive models to achieve heightened effectiveness. Building upon this perspective, Sanchez-Pinto et al. (2018) emphasized the critical importance of tailoring feature selection methods to specific dataset characteristics. Their work accentuates the need for a nuanced approach, distinguishing between classic regression-based and tree-based techniques for predictive variables, tailored to the scale of clinical datasets. Furthermore, Chowdhury and Turin (2020) highlighted the significance of variable selection in the realm of clinical prediction modeling. They underscored the potential consequences of failing to incorporate the right variables into the model, emphasizing the risk of inaccurate results could cause the model to overlook genuine relationships within the data pertaining to the outcome and the chosen variables. The discussions by Bagherzadeh-Khiabani et al. (2016) and Sanchez-Pinto et al. (2018) collectively emphasize the crucial role variable selection methods play in the optimization of clinical prediction models.

Computational efficiency

Speiser et al. (2019) emphasized the pivotal role of variable selection methods in augmenting computational efficiency within predictive modeling. Their research underscored the significance of reducing the number of predictor variables, asserting this approach is paramount for enhancing the efficiency of data collection and analysis. This aligns seamlessly with the overarching objective of elevating healthcare predictive models. Additionally, Tsamardinos et al. (2022) highlighted the transformative potential of AutoML and the JADBio platform in advancing predictive modeling and FS, particularly in precision oncology and translational medicine. This accentuated the critical importance of employing efficient and accurate FS methods to enhance predictive models in the realm of healthcare. The study conducted by Su and Yang (2008), compared the SVM-based method with the backpropagation neural network method, revealing the SVM-based approach exhibited superior performance in terms of sensitivity and specificity—crucial epidemiological indices for hypertension diagnosis.

Efficient FS methods, as explored by Speiser et al. (2019) and Tsamardinos et al. (2022), significantly contribute to the overall improvement of computational efficiency in healthcare predictive modeling. In conclusion, the literature review offers a comprehensive insight into the pivotal components shaping the domain of machine learning, including the importance of FS, the role of machine learning in healthcare, hybrid approaches, evaluation metrics, challenges in healthcare data, disease diagnosis, variable selection methods, and computational efficiency. FS techniques and predictive modeling emerge as critical elements, contributing substantially to the creation of precise, efficient, and interpretable models. These advancements extend to various applications, notably in the crucial domain of early disease prediction. The profound insights gleaned from the literature pave the way for further exploration and refinement in the research of enhanced predictive modeling within machine learning environments.

Methodology

In this study, a qualitative narrative synthesis approach was employed to analyze selected publications within the thematic systematic review. Drawing upon principles of qualitative evidence synthesis (QES) and meta-ethnography, as outlined by Bearman and Dawson (Kelly et al., n.d.), the approach aimed to extract and synthesize themes or patterns from the literature, facilitating a comprehensive understanding of the subject matter.

Booth (2016) emphasizes the importance of qualitative systematic reviews, which provide insights into how interventions in clinical trials work and their impact on different groups, informing the design of future interventions through the exploration of real-world factors.
To identify relevant literature, systematic forward and backward citation searches of key articles published over the previous 15–20 years were conducted. Google Scholar served as the primary platform for accessing titles and abstracts related to diabetes diagnostic FS keywords, relevant empirical data, and literature reviews. The inclusion criteria were determined based on the content of the abstract texts, focusing on FS strategies and predictive modeling in diabetes diagnosis as determined by the researcher. To identify relevant literature, the researcher conducted systematic searches of electronic databases including PubMed, Scopus, and Web of Science, using a combination of search terms related to diabetes diagnostic FS, predictive modeling, and qualitative evidence synthesis (QES). The search strategy included keywords such as "diabetes diagnosis," "feature selection," "predictive modeling," and "qualitative evidence synthesis," combined with Boolean operators to ensure comprehensive coverage of the literature. Searches were limited to articles published in the past 15–20 years to focus on recent advancements in the field.

Articles identified through the initial database searches underwent a two-stage screening process. In the first stage, titles and abstracts were screened for relevance to the study objectives. In the second stage, full-text articles were reviewed to determine eligibility for inclusion based on predefined criteria. The inclusion criteria were based on the relevance of the study to FS strategies and predictive modeling in diabetes diagnosis. Specifically, articles were included if they provided insights into the development, evaluation, or comparison of FS techniques with applications in predictive modeling for diabetes diagnosis. Literature reviews and empirical studies were prioritized for inclusion to ensure a comprehensive overview of the topic.

A quality assessment of included studies was conducted using the Critical Appraisal Skills Programme (CASP) tool for qualitative research (Long et al., 2020). This involved assessing the methodological rigor and credibility of the studies to determine their suitability for inclusion in the synthesis. Data extraction was performed using a standardized form to capture relevant information from included studies, including study characteristics, key findings, and methodological details. The extracted data were synthesized using thematic analysis to identify recurrent themes and patterns across the literature.

Potential biases in the literature search and selection process, such as publication bias or language bias, were acknowledged. Strategies implemented to minimize bias included comprehensive database searches, screening by multiple reviewers, and transparent reporting of the study methods.

The data analysis phase involved a qualitative systematic review aimed at synthesizing themes from the selected articles. The identified data underwent thematic analysis to assess the effectiveness of FS in enhancing diabetes diagnosis prediction modeling. This analysis aimed to provide valuable insights into the most promising techniques and their implications for improving the accuracy and reliability of predictive models.

Results

Research Question (RQ1): What are the predominant themes within the literature on FS techniques, and how do these themes influence the accuracy and reliability of predictive models in diabetes diagnosis?

Theme 1: Importance of feature selection

The literature review underscored the critical significance of FS in enhancing the accuracy and efficiency of predictive models for diabetes diagnosis. Balogun et al. (2020) conducted a comprehensive benchmark study involving 46 FS methods and 25 defect datasets, highlighting FS's pivotal role in addressing high dimensionality issues within software defect prediction (SDP). Similarly, Battineni et al. (2020) emphasized the importance of selecting pertinent features in machine learning for chronic illness diagnosis, emphasizing FS's role in achieving precision within predictive modeling. These findings reaffirm the foundational importance of FS techniques in improving the accuracy and reliability of predictive models, particularly in diabetes diagnosis.

Theme 2: Integration of machine learning in healthcare

The integration of machine learning algorithms with FS techniques emerged as a promising approach for advancing predictive modeling in healthcare, including diabetes diagnosis. Kanyongo and Ezugwu (2023) demonstrated the indispensable role played by machine learning in predicting medication adherence for non-communicable diseases, showcasing the transformative impact of machine learning techniques in healthcare domains. Spencer et al. (2020) emphasized the advantages of uniting FS with machine learning algorithms to predict heart disease, further highlighting machine learning's potential in enhancing diagnostic accuracy. These findings underscore the growing importance of machine learning in healthcare and its potential to revolutionize predictive modeling for disease diagnosis, including diabetes.

Theme 3: Exploration of hybrid approaches

Researchers increasingly advocate for hybrid classification approaches to enhance classifier accuracy and computational efficiency in predictive modeling. Jain and Singh (2018) systematically categorized FS algorithms and advocated for hybrid classification approaches, aiming to improve predictive models' accuracy, particularly for chronic disease prediction. Rauber et al. (2015) demonstrated the effectiveness of their strategy involving the integration of various feature models within a unified pool, highlighting the potential of hybrid approaches to optimize predictive modeling for diabetes diagnosis. These studies offer promising avenues for future research in enhancing predictive modeling techniques for diabetes diagnosis.

Theme 4: Evaluation metrics in healthcare

The evaluation of predictive models in healthcare necessitates appropriate metrics to assess their effectiveness. Mohammad et al. (2022) emphasized the crucial role of performance indicators such as accuracy, recall, and F1-measure in evaluating software fault prediction models. Olivera et al. (2017) advocated for the utilization of the area under the curve (AUC) to assess the predictive effectiveness in diagnosing undiagnosed diabetes. These findings underscore the importance of selecting appropriate evaluation metrics to ensure the reliability and validity of predictive models for diabetes diagnosis.

Theme 5: Challenges and opportunities in healthcare data analysis

Addressing challenges associated with healthcare data analysis is crucial for developing accurate and reliable predictive models for diabetes diagnosis. Patiño-Saucedo et al. (2022) highlighted the challenges associated with domain-specific FS and optimization difficulties crucial for accurate healthcare diagnosis. Reps et al. (2018) introduced a structured framework for developing patient-level prediction models utilizing observational healthcare data, addressing the need for transparent and reproducible approaches in healthcare data analysis. These findings highlight significant challenges and opportunities in healthcare data analysis and underscore the importance of adopting rigorous methodologies in predictive modeling research.

Discussion of findings

The systematic review undertaken in this study illuminates the crucial role of feature selection (FS) techniques in enhancing predictive modeling for diabetes diagnosis. Through an exhaustive synthesis of literature on machine learning applications in healthcare, several key themes emerged, shedding light on significant implications for both research and clinical practice.

Implications of findings

The review underscores the importance of FS as a foundational component in refining predictive models for diabetes diagnosis. Notably, FS techniques address high dimensionality issues, improve model accuracy, and enhance computational efficiency. This highlights the necessity for researchers and practitioners to prioritize the selection and optimization of FS techniques in model development.

Integration of machine learning algorithms

Moreover, integrating machine learning algorithms with FS techniques offers a promising avenue for advancing predictive modeling in healthcare, especially in diabetes diagnosis. The transformative impact of machine learning on enhancing diagnostic accuracy and efficiency is evident from the reviewed studies. Leveraging machine learning alongside FS techniques enables the development of robust predictive models capable of handling complex healthcare data.

Recommendations for future research

Exploring hybrid classification approaches, which combine multiple FS techniques and machine learning algorithms, presents potential synergies for further enhancing predictive model accuracy and computational efficiency. By integrating diverse methodologies, researchers can develop innovative solutions tailored to the specific challenges of diabetes diagnosis.

Evaluation metrics and limitations

Selecting appropriate evaluation metrics, such as accuracy, recall, and area under the curve (AUC), is crucial for assessing the effectiveness of predictive models in healthcare. Future research must carefully consider the choice of evaluation metrics to accurately assess model performance and guide clinical decision-making. Additionally, addressing challenges in healthcare data analysis, including domain-specific FS issues and optimization difficulties, requires interdisciplinary collaboration and rigorous methodologies.

Conclusion

In conclusion, this systematic review provides valuable insights into the role of FS techniques in predictive modeling for diabetes diagnosis. By synthesizing existing literature and identifying key themes, this study informs researchers, practitioners, and policymakers involved in diabetes care. Moving forward, future research should focus on developing innovative FS methodologies, integrating machine learning algorithms, and addressing challenges in healthcare data analysis to advance predictive modeling and enhance patient outcomes.

References

Alhassan, A. M., & Wan Zainon, W. M. (2021). Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis. *IEEE Access, 9*, 87310–87317. https://doi.org/10.1109/access.2021.3088613

Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., & Khalili, D. (2016). A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results. *Journal of Clinical Epidemiology, 71*, 76–85. https://doi.org/10.1016/j.jclinepi.2015.10.002

Balogun, A. O., Basri, S., Mahamad, S., Abdulkadir, S. J., Almomani, M. A., Adeyemo, V. E., ... Bajeh, A. O. (2020, July 9). Impact of feature selection methods on the predictive performance of software defect prediction models: An extensive empirical study. *Symmetry, 12*(7), 1147. https://doi.org/10.3390/sym12071147

Bashir, S., Khan, Z. S., Hassan Khan, F., Anjum, A., & Bashir, K. (2019). Improving heart disease prediction using feature selection approaches. *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. https://doi.org/10.1109/ibcast.2019.8667106

Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020, March 31). Applications of machine learning predictive models in chronic disease diagnosis. *Journal of Personalized Medicine, 10*(2), 21. https://doi.org/10.3390/jpm10020021

Booth, A. (2016). Searching for qualitative research for inclusion in systematic reviews: A structured methodological review. *Systematic Reviews, 5*, 74. https://doi.org/10.1186/s13643-016-0249-x

Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. (2021). A multi-stage approach combining feature selection with machine learning techniques for higher prediction reliability and accuracy in cervical cancer diagnosis. *International Journal of Intelligent Systems and Applications, 13*(5), 46–63. https://doi.org/10.5815/ijisa.2021.05.05

Chowdhury, M. Z. I., & Turin, T. C. (2020, February). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health, 8*(1), e000262. https://doi.org/10.1136/fmch-2019-000262

Ershadi, M. M., & Seifi, A. (2022). Applications of dynamic feature selection and clustering methods to medical diagnosis. *Applied Soft Computing, 126*, 109293. https://doi.org/10.1016/j.asoc.2022.109293

Hassan, S., Dhali, M., Zaman, F., & Tanveer, M. (2021). Big data and predictive analytics in healthcare in Bangladesh: regulatory challenges. *Heliyon, 7*(6).

Jain, D., & Singh, V. (2018, November). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal, 19*(3), 179–189. https://doi.org/10.1016/j.eij.2018.03.002

Kanyongo, W., & Ezugwu, A. E. (2023). Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. *Informatics in Medicine Unlocked, 38*, 101232. https://doi.org/10.1016/j.imu.2023.101232

Kelly, M., Reid, H., Bennett, D., Yardley, S., & Dornan, T. (n.d.). Introduction to qualitative evidence synthesis. *The Arnold P. Gold Foundation*. Retrieved January 30, 2024, from https://www.gold-foundation.org/programs/research/mtl/introduction-qualitative-evidence-synthesis/#:~:text=Noblit%20and%20Hare%20%E2%80%99s%20work%20was%20seminal%20because,first%20time%20a%20way%20to%20synthesize%20qualitative%20studies

Long, H., French, D., & Brooks, J. (2020). Optimising the value of the Critical Appraisal Skills Programme (CASP) tool for quality appraisal in qualitative evidence synthesis. *Research Methods in Medicine and Health Sciences, 1*(1), 31–42. https://doi.org/10.1177/2632084320947559

Mohammad, U. G., Imtiaz, S., Shakya, M., Almadhor, A., & Anwar, F. (2022, June 27). An optimized feature selection method using ensemble classifiers in software defect prediction for healthcare systems. *Wireless Communications and Mobile Computing, 2022*, 1–14. https://doi.org/10.1155/2022/1028175

Olivera, A. R., Roesler, V., Iochpe, C., Schmidt, M. I., Vigo, L., Barreto, S. M., & Duncan, B. B. (2017, June). Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: Accuracy study. *Sao Paulo Medical Journal, 135*(3), 234–246. https://doi.org/10.1590/1516-3180.2016.0309010217

Patiño-Saucedo, J. A., Ariza-Colpas, P. P., Butt-Aziz, S., Piñeres-Melo, M. A., López-Ruiz, J. L., Morales-Ortega, R. C., ... De-la-hoz-Franco, E. (2022). Predictive model for human activity recognition based on machine learning and feature selection techniques. *International Journal of Environmental Research and Public Health, 19*(19), 12272. https://doi.org/10.3390/ijerph191912272

Rauber, T. W., de Assis Boldt, F., & Varejao, F. M. (2015). Heterogeneous feature models and feature selection applied to Bearing Fault Diagnosis. *IEEE Transactions on Industrial Electronics, 62*(1), 637–646. https://doi.org/10.1109/tie.2014.2327589

Reddy, N. S. C., Nee, S. S., Min, L. Z., & Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. *International Journal of Innovative Computing, 9*(1), 210. https://doi.org/10.11113/ijic.v9n1.210

Reps, J. M., Schuemie, M. J., Suchard, M. A., Ryan, P. B., & Rijnbeek, P. R. (2018). Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association, 25*(8), 969–975. https://doi.org/10.1093/jamia/ocy032

Sakri, S. B., Abdul Rashid, N. B., & Muhammad Zain, Z. (2018). Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction. *IEEE Access, 6*, 29637-29647. https://doi.org/10.1109/ACCESS.2018.2843443

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., & Churpek, M. M. (2018). Comparison of variable selection methods for clinical predictive modeling. *International Journal of Medical Informatics, 116*, 10–17. https://doi.org/10.1016/j.ijmedinf.2018.05.006

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications, 40*(10), 4146–4153. https://doi.org/10.1016/j.eswa.2013.01.032

Shreffler, J. (2023, March 3). Diagnostic testing accuracy: Sensitivity, specificity, predictive... *National Library of Medicine*. https://www.ncbi.nlm.nih.gov/books/NBK557491/

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications, 134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020, January). Exploring feature selection and classification methods for predicting heart disease. *DIGITAL HEALTH, 6*, 205520762091477. https://doi.org/10.1177/2055207620914777

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. Advance online publication. https://doi.org/10.1002/widm.1379

Su, C. T., & Yang, C. H. (2008). Feature selection for the SVM: An application to hypertension diagnosis. *Expert Systems with Applications, 34*(1), 754–763. https://doi.org/10.1016/j.eswa.2006.10.010

Tsamardinos, I., Charonyktakis, P., Papoutsoglou, G., Borboudakis, G., Lakiotaki, K., Zenklusen, J. C., ... Lagani, V. (2022, June 16). Just add data: Automated predictive modeling for knowledge discovery and feature selection. *Npj Precision Oncology, 6*(1). https://doi.org/10.1038/s41698-022-00274-8

World Health Organization. (n.d.). Diabetes. *World Health Organization*. Retrieved November 2, 2023, from https://www.who.int/health-topics/diabetes#tab=tab_1

Wu, J., Roy, J., & Stewart, W. F. (2010, June). Prediction modeling using EHR data. *Medical Care, 48*(6), S106–S113. https://doi.org/10.1097/mlr.0b013e3181de9e17