Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241363

Exploring the current status and application of machine learning in disease risk prediction models in medical examination

Shengyu Xiao

University of Phoenix, No.48, Binhu Road, Nanning, Guangxi, 530022, China

xiaos3@email.phoenix.edu

Abstract. As the aging population problem continues to grow in Asia, an efficient medical examination system based on machine learning will be required. By detecting diseases early, healthcare professionals can provide early interventions and treatments to improve patient's quality of life and overall health outcomes. Therefore, this paper explores the feasibility of utilizing disease risk prediction models based on lifestyle factors in medical examinations and examines possible implications that could arise from applying such models. This study aims to discuss the potential application of machine learning in disease risk prediction models based on lifestyle factors, such as heart disease and cancer, in medical examination. The aim is to use machine learning models in medical examinations to detect diseases in their early stages. The lifestyle factors can be collected through forms and put into a general machine learning model. There will be three machine learning models that could be used as references, which could be used to support the idea's feasibility.

Keywords: Machine Learning, Prediction Model, Medical Examination, Lifestyle factors.

1. Introduction

The escalating issue of an aging population in Asia necessitates implementing an effective health assessment system that utilizes machine learning techniques. The aging rate in China has steadily increased by approximately 3.2% per year since 1970 [1]. Based on the findings of the seventh national census, it was observed that the elderly population in China reached a significant figure of 264 million individuals in 2020 [2]. This demographic trend implies a corresponding increase in the prevalence of age-related ailments, including chronic diseases, malignancies, and cardiovascular disorders. Furthermore, with the evolving nature of lifestyles, many young individuals are experiencing suboptimal health conditions, potentially leading to the development of diverse ailments in the absence of regular annual check-ups.

The timely identification of diseases can enhance therapy efficacy and prognosis. Therefore, it is crucial to enhance the dissemination of knowledge regarding the significance of routine health screenings due to the escalating prevalence of diverse health ailments. The timely identification of disorders enables healthcare practitioners to administer prompt interventions and therapies, enhancing patients' quality of life and overall health outcomes. It is anticipated that there will be a growing demand for health assessments in the forthcoming decades, owing to the ongoing expansion and aging of the

DOI: 10.54254/2755-2721/46/20241363

population. This study addresses the potential enhancement of early illness detection efficiency by integrating disease risk prediction models based on lifestyle factors with frequent check-ups. The subsequent passages will delve into this topic.

The illness risk prediction model is gaining significance as the field of machine learning undergoes continuous expansion and evolution. Due to its capacity to effectively forecast the probability of acquiring specific ailments, this technology possesses the potential to revolutionize the existing paradigm of health assessment, ultimately enhancing the healthcare sector. One of the most auspicious applications of this technology pertains to health assessment. Through machine learning, healthcare practitioners can effectively harness its capabilities to better understand a patient's well-being and proactively detect potential concerns before they escalate into significant complications [3]. This kind of action can result in enhanced treatment programs tailored to individual patients, ultimately leading to improved patient outcomes and overall quality of life.

Based on the provided facts, it is evident that the development of a machine learning model incorporating lifestyle characteristics as data holds the potential for enhancing medical examination practices by enabling the provision of individualized medical examination plans. This model can facilitate the identification of diseases during their initial phases. Therefore, the objective of this study is to investigate the viability of employing disease risk prediction models founded on lifestyle factors during health assessments, as well as the potential consequences that may develop as a result of implementing these models. This work examines the possible utilization of machine learning techniques in disease risk prediction models incorporating lifestyle factors, including heart disease and cancer, within medical examinations. The objective is to employ machine learning models to identify diseases at their initial phases in medical screenings. The collection of lifestyle factors can be facilitated by utilizing forms, which can then be inputted into a comprehensive machine-learning model. The proposed algorithm will recommend that patients undergo additional examinations depending on the projected outcomes. Three machine learning models will be shown as references to support the practicality of the proposed proposal [4].

2. Literature Review

2.1. Algorithm modeling strategies

The CanICU model, developed by Ko, uses machine learning techniques to estimate the probability of mortality during the first 28 days for adult cancer patients admitted to the intensive care unit (ICU) [4]. The model was designed specifically for the healthcare industry and incorporates a range of patient characteristics that are relevant to the field. This comprehensive approach allows the model to achieve high sensitivity and perform effectively across different validation cohorts. The model demonstrates notable sensitivity and specificity since it incorporates nine factors. Consequently, it exhibits enhanced predictive capabilities for both short- and long-term mortality outcomes. The CanICU system assists physicians in allocating intensive care unit (ICU) resources for cancer patients, utilizing an objective mortality risk assessment.

The CanICU report includes a section that delineates the development of a prognostic model. The cohort of patients in the YCC study was partitioned into two groups: a derivation cohort including 70% of the patients, and an internal validation cohort comprising the remaining 30% of patients. The development of CanICU involved the utilization of three distinct modelling methodologies, namely random forest, extreme gradient boosting (XGBoost), and support vector machine (SVM). To achieve the class balance between the key outcome groups, the derivation cohort employed the synthetic minority oversampling method (SMOTE) to oversample the minority class (fatal cases within 28 days) by a factor of five. The random forest algorithm demonstrated the highest level of sensitivity and was subsequently reported due to its robustness. The 'CanICU' model was constructed exclusively with patients who exhibited all nine predetermined variables. According to Ko, the ultimate forecast is established based on the predominant consensus of votes obtained from numerous decision trees [4]. The efficacy of the CanICU model in mortality prediction for adult cancer patients hospitalized in the

DOI: 10.54254/2755-2721/46/20241363

intensive care unit (ICU) has been well-established. The advancement mentioned above holds significant implications within the realm of cancer care since the ability to forecast mortality enables healthcare practitioners to make well-informed judgments regarding the course of therapy for their patients. Furthermore, using random forest modelling techniques and dataset validation methodologies employed in constructing the CanICU model can serve as a valuable framework for making other machine learning models within the healthcare domain [5]. This will facilitate the endeavors of researchers and developers in building precise and practical models that can aid in advancing machine learning models for health assessment.

2.2. Algorithm Data Selection

The research conducted by Afrash in the paper titled "Development of machine learning models for early prediction of gastric cancer risk using lifestyle factors" employed machine learning methodologies to create predictive models for estimating the likelihood of developing gastric cancer [6]. The models were constructed using data obtained from a comprehensive gastric cancer database [6]. The research identified eleven significant characteristics that influence susceptibility to gastric cancer, with the XGBoost model demonstrating superior efficacy in forecasting the risk of gastric cancer. The findings indicate that the application of machine learning methods holds promise in the preliminary screening of stomach cancer and identifying individuals at heightened risk, leading to a substantial decrease in the number of patients necessitating endoscopic surveillance.

The study offers a comprehensive elucidation of the data preprocessing procedure within the context of disease risk prediction models in health assessment, employing the CRISP-DM approach and machine learning techniques. The main aim of data preprocessing is to appropriately prepare the data for analysis by eliminating outliers, imputing missing values, and transforming the data into a format conducive to trustworthy and accurate analysis. The essay highlights the significance of data preparation within the context of machine learning and its consequential effects on data mining methodologies.

In the initial data preprocessing stage, the interquartile range rule is employed to identify and eliminate outliers from the dataset. Additionally, the dataset is normalized using the min-max approach. The article also elucidates the mean and regression-based techniques employed for imputing missing values and deleting rows containing over 70% missing values.

Data scaling is a crucial component of data preprocessing, as elucidated in the article. The paper delineates two distinct approaches employed for data scaling: data distribution-based scaling and data range-based scaling. The Z-score standardization is utilized as a scaling approach based on the data distribution, whereas the min-max method is employed for scaling based on the range of data. These methodologies guarantee data normalization on the same scale, enabling the model to successfully acquire knowledge from the data.

The essay additionally emphasizes the concern of unbalanced datasets and their influence on the efficacy of machine learning models. A new dataset is generated by approximating the group with a smaller number of samples to the group with a more significant number of instances, as observed in the dataset. The application of the synthetic minority oversampling method (SMOTE) involves the process of resampling the data to achieve a balanced dataset. The paper elucidates that SMOTE is widely used as the predominant and productive technique for oversampling in diverse domains to achieve dataset balance.

Upon doing a comprehensive analysis of the paper, a more profound comprehension was attained regarding the use of the CRISP-DM approach and SMOTE technique to augment the precision of the dataset. The essay further emphasizes the significance of a dependable and all-encompassing dataset to construct a more accurate model. In addition, the study examines the CRISP-DM approach, a widely acknowledged framework within the domain of data mining. This methodology encompasses six distinct phases: business understanding, data understanding, data preparation, modeling, assessment, and deployment. The SMOTE technique, conversely, is a methodology employed to mitigate the problem of class imbalance within the dataset. SMOTE, by the technique of oversampling the minority class, contributes to mitigating bias towards the majority class within the model, enhancing its accuracy and

DOI: 10.54254/2755-2721/46/20241363

dependability. The essay offers insights into the optimal approaches for constructing an accurate model by employing various strategies and processes.

Moreover, Afrash includes various lifestyle factors such as age, gender, BMI, blood type, marital status, family history of cancer, depression, stress status, income level, education level, residence status, high salt intake, high-fat foods status, alcohol consumption, smoking, physical activity, fruit intake, red meat consumption, gastric cancer screening participation, weight loss, Helicobacter pylori test results, upper abdominal fullness, abdominal pain, stomach polyps, recurrent nausea and vomiting, stomach or duodenal ulcers, and previous stomach surgery or chronic atrophic gastritis [6]. Using data types such as stress status, income level, high salt intake, and alcohol consumption can significantly enhance the effectiveness of a medical examination model developed in the future. These data types are closely associated with various diseases. However, certain data types, such as stomach polyps, vomiting, and gastric cancer screening participation, which are specific to stomach cancer, may need to be excluded from a general medical examination aiding model to ensure optimal efficiency.

Therefore, it is essential to establish a comprehensive lifestyle factor dataset that connects various diseases. By building such a dataset, we can create a more holistic medical examination model that captures a broader range of factors affecting overall health and well-being. This will allow healthcare providers to identify the potential risks of diseases better and provide more targeted and effective treatment plans. So, it is crucial to focus on collecting and analyzing lifestyle data to improve the quality and accuracy of medical examination models.

2.3. An Analogous Approach in Machine Learning

Effiok presents a risk prediction framework for assessing the cumulative impact of lifestyle risk variables on complex diseases [7]. The framework is applied to case studies to demonstrate its efficacy. The aim is to propose a foundational framework for forecasting favorable and unfavorable outcomes in intricate medical conditions. This framework can serve as a repository of information for Internet of Things (IoT)-enabled healthcare applications and facilitate tailored risk assessment for patients' self-care, coaching services, and patient education.

This work compares two distinct machine learning algorithms, namely GCALC and SWOP, in the context of illness risk prediction during Medical Examinations. The study is centered around the assessment of the algorithms' performance through the utilization of a cross-validation technique. In this approach, both algorithms underwent testing utilizing identical patient data.

The algorithms underwent evaluation using a cohort of five patients, all falling within the age bracket of 70 to 74 years and lacking any familial predisposition to prostate cancer. Every individual in the study population had experienced one or more urinary conditions and had also been subjected to lifestyle risk factors that are recognized to impact the development of prostate cancer. The evaluation data for each patient encompassed distinct lifestyle risk variables and urine issue data from the initial questionnaires administered to each patient. The data was utilized to forecast the likelihood of disease incidence by applying GCALC and SWOP algorithms.

The findings of the study indicated that, in general, the GCALC algorithm tended to yield higher predictions of illness incidence when compared to the SWOP algorithm. However, it is worth noting that there was one instance in which both algorithms produced low probabilities of disease occurrence for a particular patient. As an illustration, Patient 1, a male individual aged 70 with obesity, regular aspirin usage, and urinary complications, exhibited a 60% probability of illness manifestation when assessed using GCALC, but the likelihood was determined to be 12% when employing SWOP. In line with this, Patient 5, a 71-year-old male who is obese, regularly uses aspirin, and has been exposed to arsenic, exhibited a 68% probability of illness manifestation when assessed using the GCALC method. In comparison, the SWOP method yielded a 12% likelihood.

The research also emphasizes the significance of conducting comparative analyses of many predictive models that are designed to perform comparable tasks. Such comparisons facilitate the identification of fresh insights for critical analysis and potentially result in enhancements to one or a few of the models under consideration. In general, the study offers significant insights into using machine

DOI: 10.54254/2755-2721/46/20241363

learning algorithms for predicting illness risk in medical examinations, underscoring the necessity for additional research in this domain to enhance the precision and dependability of said algorithms.

3. Methodology

This methodology provides an overview of the components of building machine learning models, including data collection and preprocessing, model selection and development, and model evaluation and interpretation. A potential application scenario is presented to demonstrate how this methodology can be applied in real-world situations.

Collecting and preprocessing data are crucial when building a machine-learning model. The data quality in the model directly affects its accuracy and reliability. Therefore, it is essential to use various methods for data collection, including surveys, electronic health records, and wearable technology [8]. After data collection, it is necessary to undergo preprocessing to eliminate outliers, eliminate noisy data, fill in missing values, and transform the data into a format appropriate for a thorough and precise analysis.

In the early stages of model development, various datasets can be used. However, it is crucial to establish a standard for which lifestyle factors are essential. These factors should be easy to collect, analyze, and connect to various diseases. This will ensure the data is relevant, accurate, and valuable in developing the model.

The next model development phase can begin once the data has been collected and preprocessed. Although the CanICU machine learning model uses random forest trees as its primary algorithm, various algorithms can be tried to see if they fit this program. The process of model development includes selecting and developing a suitable machine-learning model. Machine learning algorithms can be tested if they perform better than random forests, such as decision trees, support vector machines, linear regression [5]. The model development process includes training, validating, and testing the model to ensure its accuracy and reliability.

After the model has been developed, it needs to be evaluated and interpreted to determine its accuracy and reliability. This evaluation is critical to ensure the model performs well in real-world scenarios. Several evaluation metrics can be used to evaluate the model's performance, for example, the cohort mentioned in CanICU. It is essential to carefully select the most appropriate metrics based on the specific problem and data at hand.

Once the evaluation process is completed, the model's results can be interpreted to gain insights into its behavior and performance. This interpretation process can reveal the essential features contributing to the model's predictions and the decision boundaries separating different classes. By understanding these aspects of the model, we can gain a deeper understanding of the problem being solved and the model's strengths and weaknesses. This knowledge can then be used to improve the model and its performance.

In summary, collecting and preprocessing data, selecting and developing a suitable machine learning model, and evaluating and interpreting the model are all crucial steps in building a machine learning model. These steps must be carried out carefully and systematically to ensure the model is accurate, reliable, and effective. By following these steps, we can create machine learning models that can assist in developing medical examination models and improve healthcare outcomes.

4. Discussion and Future Application Scenarios

One potential application of machine learning in healthcare is to integrate a machine learning model into a software application, whether for a phone or desktop, to assist patients during their annual health check-ups. The risk model can predict various diseases based on lifestyle factors by completing a short survey. Once the risk model has made predictions, it can offer a more specific and customized medical examination plan, increasing the possibility of detecting diseases such as cancer more efficiently. Early detection of conditions can lead to better treatment outcomes and improved prognosis.

Considering the current aging situation and high-stress levels in healthcare, this application looks promising. There could be a rise in the need for medical examination resources, and such a machine

DOI: 10.54254/2755-2721/46/20241363

learning model would allow for much better utilization of resources and save a great deal of money in social and medical insurance.

However, there are still concerns regarding collecting personal data and the potential loss of privacy and autonomy in utilizing machine learning models in healthcare. Patients may not be willing to provide more detailed lifestyle factors if certain requirements are not met, which could impede the effectiveness of such models [9]. Furthermore, any application that collects and utilizes personal data must have a secure server to ensure the privacy of users' information.

One potential solution to this issue is to connect the machine learning model with the hospital information system (HIS). However, due to the varying HISs used by hospitals, there may be challenges in modifying the application of the model. These challenges could include difficulties in data extraction and integration and issues with standardization across different HIS [10].

Therefore, it is essential to carefully consider the feasibility and practicality of utilizing machine learning models based on lifestyle factors in medical examination, particularly regarding data collection and integration. It is essential to prioritize patients' privacy and autonomy while ensuring the models' accuracy and effectiveness. This can be achieved by establishing secure servers and developing standardized data collection methods that can be easily integrated with different hospital information systems. Addressing these concerns can lead to developing machine learning models to help healthcare professionals in early disease detection and personalized treatment plans, ultimately improving patient outcomes and quality of life.

5. Conclusion

This paper delves into the exciting possibilities of applying machine learning to disease risk prediction models based on lifestyle factors in medical examinations. It initially introduces the concept of disease risk prediction models and the potential benefits they hold for improving healthcare outcomes. The paper then describes three machine learning models that could be utilized in developing such models, providing a detailed explanation of each model's strengths and weaknesses.

In addition to discussing these models, the paper proposes a potential methodology for building a general risk prediction model in Medical Examinations, outlining the steps that need to be taken to ensure its successful implementation. It also explores the feasibility of integrating machine learning models into software applications to assist patients during their annual health check-ups, suggesting that such an approach could help improve patient outcomes and reduce healthcare costs.

Overall, this paper provides valuable insights into the potential benefits of utilizing machine learning in healthcare, offering a detailed and comprehensive exploration of the topic that should interest healthcare professionals, researchers, and policymakers alike.

References

[1] Bao, J., Zhou, L., Liu, G., Tang, J., Lu, X., Cheng, C., Jin, Y., and Bai, J 2022 Current state of care for the elderly in China in the context of an aging population *BioScie. Tre.* 16(2) 107-118

[2] Ma, C., Xu, W., Zhou, L., Ma, S., and Wang, Y 2018 Association between lifestyle factors and suboptimal health status among chinese college freshmen: A cross-sectional study *BMC Public Health* 18(1)

[3] Garg, A., and Mago, V. 2021 Role of machine learning in Medical Research: A survey *Comp. Sci. Revi.* 40 100370

[4] Ko, R.-E., Cho, J., Shin, M.-K., Oh, S. W., Seong, Y., Jeon, J., Jeon, K., Paik, S., Lim, J. S., Shin, S. J., Ahn, J. B., Park, J. H., You, S. C., and Kim, H. S 2023 Machine learning-based mortality prediction model for critically ill cancer patients admitted to the Intensive Care Unit (canicu) *Cancers* 15(3) 569

[5] Alpaydin, E. 2020 The MIT Press Introduction to machine learning

[6] Afrash, M. R., Shafiee, M., and Kazemi-Arpanahi, H 2023 Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors *BMC Gastro.* 23(1)

[7] Effiok, E., Liu, E., and Hitchcock, J. 2022 Health. Tech. Letters 9(3) 34–42

DOI: 10.54254/2755-2721/46/20241363

[8] Pandey, S. C. 2016 Predicting cumulative effect of lifestyle risk factors for complex disease SCOPES Data mining techniques for Medical Data: A Review *Inter. Conf. on Signal Proc.*

[9] Köngeter, A., Schickhardt, C., Jungkunz, M., Bergbold, S., Mehlis, K., and Winkler, E. C. 2022 Patients' willingness to provide their clinical data for research purposes and acceptance of different consent models: Findings from a representative survey of patients with cancer *Jour. of Med. Inter. Res.* 24(8)

[10] RHIA, N. E. M., George, H. T., Arias, J. J., and Sompalli, R. 2017 Hospital Information Systems: Electronic Medical Record, EMR software, electronic health record, EHR software customized recommendations *MPM* 33(3) 184-186