Educational Administration: Theory and Practice
2024, 30(5), 13647 - 13651
ISSN: 2148-2403
https://kuey.net/

Research Article

Leveraging Support Vector Machines For Early Disease Detection Using Electronic Health Records

Dr. S. Uma[1*]

[1*] Associate Professor in Computer Science, Dr.N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India

Citation: Dr. S. Uma (2024) Leveraging Support Vector Machines For Early Disease Detection Using Electronic Health Records, *Educational Administration: Theory and Practice, 30(5),* 13647 - 13651
Doi: 10.53555/kuey.v30i5.5935

ARTICLE INFO **ABSTRACT**

Early disease detection using electronic health records (EHRs) is vital for improving patient outcomes and reducing healthcare costs. This proposed methodology integrates Support Vector Machines (SVM) for predictive modeling, emphasizing data preprocessing, feature selection, and SVM model design. Key steps include data collection, preprocessing, feature selection, and SVM model training with various kernel functions. The decision function of the SVM is described, and the SVM with kernel algorithm is outlined. This approach aims to enhance early disease detection capabilities, leveraging EHR data effectively.

Keywords: Early disease detection, Electronic health records (EHRs), Support Vector Machines (SVM), Predictive modeling, Data mining.

1. Introduction

Predictive modeling for early disease detection has become a cornerstone of modern healthcare, offering the potential to transform patient outcomes through timely intervention. As the volume of electronic health records (EHRs) continues to grow, harnessing this wealth of data through advanced data mining techniques has emerged as a critical strategy for healthcare providers. EHRs, which compile comprehensive patient information including demographics, medical history, laboratory results, medications, and clinical notes,offer a wealth of information for creating prediction models that can spot disease symptoms early.Early detection of diseases significantly improves treatment success rates and reduces healthcare costs by enabling preventive care and early intervention. Traditional diagnostic methods often rely on clinical symptoms that manifest in the later stages of a disease, thus limiting the window for effective intervention. Predictive modeling, however, leverages historical and real-time data from EHRs to identify patterns and indicators that precede the onset of symptoms, offering a proactive approach to disease management.

Predictive modeling heavily relies on data mining, the process of finding patterns and relationships in huge datasets. Techniques such as classification, regression, clustering, and anomaly detection are employed to analyze EHR data and predict disease onset. Machine learning algorithms, including decision trees, support vector machines, neural networks, and ensemble methods, are particularly effective in handling the complexity and heterogeneity of EHR data. These algorithms can learn from vast amounts of data, making sense of intricate relationships and providing accurate predictions.One of the primary challenges in predictive modeling using EHRs is data quality. EHRs often contain missing, inconsistent, or erroneous entries, which can hinder model performance. Effective data preprocessing techniques, such as imputation for missing values, normalization, and noise reduction, are essential to ensure data quality. Additionally, the integration of various data sources, such as imaging, genomic, and wearable device data, can enrich the EHR dataset, offering a more comprehensive view of patient health.

Ethical considerations and patient privacy are paramount in the utilization of EHR data for predictive modeling. Ensuring data security and patient confidentiality through de-identification, secure data storage, and compliance with regulations such as the Health Insurance Portability. Additionally, transparency in model development and validation processes helps build trust among patients and healthcare providers, fostering the adoption of predictive modeling in clinical practice.The integration of predictive modeling into healthcare workflows requires interdisciplinary collaboration among data scientists, clinicians, and healthcare administrators. Developing user-friendly interfaces and decision support systems that seamlessly integrate predictive insights into clinical practice is crucial for effective implementation. Training healthcare professionals to interpret and act upon predictive model outputs is also necessary to maximize the benefits of these advanced technologies.

2. LiteratureSurvey

2.1Generalized Linear Models (GLMs)

Estiri H et.al proposed Predicting COVID-19 mortality with electronic medical records. This study utilizes electronic health records (EHRs) to predict post-COVID-19 mortality and discern age-specific risk factors. Using component-wise gradient boosting to train age-stratified generalised linear models (GLMs), we were able to identify 46 clinical disorders that may be associated with an increased risk of death following a COVID-19 infection. Surprisingly, the models, relying solely on pre-existing demographic and medical data, exhibited comparable performance to more complex prognostic models. Age emerged as the most critical predictor of mortality, while conditions like pneumonia, diabetes with complications, and certain cancers significantly influenced mortality risk across different age groups. These findings underscore the importance of leveraging EHR data for targeted resource allocation and vaccination prioritization.

2.2 Artificial intelligence

Goh KH et.al proposed Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. The SERA algorithm demonstrates promising advancements in early sepsis prediction and diagnosis, leveraging both structured data and unstructured clinical notes. Twelve hours before sepsis start; it performs better than doctor estimations with good predictive accuracy (AUC 0.94, sensitivity 0.87, specificity 0.87). By incorporating unstructured clinical notes, SERA enhances accuracy, showcasing its potential to increase early sepsis detection by up to 32% and reduce false positives by up to 17%. This innovative approach addresses the critical challenge of early sepsis identification, offering a significant advancement in hospital care and potentially saving numerous lives.

2.3 BEHRT: A Deep Neural Sequence Transduction Model

Li Y, Rao S. et.al proposed BEHRT: transformer for electronic health records. BEHRT exhibits remarkable advancements in early disease detection by simultaneously predicting the likelihood of 301 conditions in future patient visits. Comparing BEHRT to current deep EHR models, there is a notable improvement of 8.0–13.2% in average precision scores. BEHRT was trained and assessed on data from 1.6 million individuals.Beyond its scalability and accuracy, BEHRT offers personalized interpretation of predictions and accommodates diverse healthcare concepts, promising enhanced accuracy and potential applications in transfer learning for future studies.

3. Proposed Methodology

Early detection of diseases significantly enhances treatment efficacy, improves patient outcomes, and reduces healthcare costs. Leveraging electronic health records (EHRs) for predictive modeling offers an opportunity to identify disease patterns early through data mining techniques. This methodology proposes using Support Vector Machines (SVM) for predictive modeling in early disease detection, emphasizing a structured approach to preprocess, model, and validate EHR data.

Data Collection and Preprocessing

The first step involves gathering comprehensive EHR data from various healthcare institutions, ensuring a wide range of patient demographics, medical histories, and diagnostic outcomes. Given the sensitive nature of EHRs, all data collection must comply with relevant privacy regulations such as HIPAA. Preprocessing the EHR data is crucial to ensure its quality and relevance. This involves cleaning the data by removing duplicates, handling missing values through imputation or exclusion, and standardizing data formats. The data is then normalized or scaled to ensure uniformity, especially since SVMs are sensitive to the scale of input features.

Feature Selection and Extraction

Effective feature selection is vital for improving the model’s accuracy and reducing computational complexity. This process begins with domain knowledge to identify relevant features related to the disease of interest, such as patient age, gender, medical history, lab test results, and genetic information. Techniques like Recursive Feature Elimination (RFE) and principal component analysis (PCA) can be employed to further refine the feature set. Additionally, feature extraction from unstructured data in EHRs, such as clinical notes, can be performed using natural language processing (NLP) techniques to derive meaningful attributes.

3.1 Proposed Support Vector Machines (SVM) Model

Support Vector Machines (SVM) is selected for their robustness in handling high-dimensional data and their effectiveness in binary classification tasks. The SVM model is designed to maximize the margin between different classes in the dataset. Given the nature of early disease detection, the focus is on classifying patients into 'disease' or 'no disease' categories. The data is transformed into a higher-dimensional space using the

kernel trick, where it becomes linearly separable. Various kernel functions (linear, polynomial, radial basis function (RBF)) are evaluated to determine the optimal one for the dataset.
The decision function of an SVM can be represented as:

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) + b)$$

Where $f(x)$represents the decision function, α_i denotes Lagrange multipliers, $y_i$denote class labels, $K(x_i, x)$ is the kernel function, and b signifies the bias term.
The choice of kernel function plays a crucial role in SVM performance. Different types of kernel functions, including linear, polynomial, and radial basis function (RBF), are evaluated to determine the optimal one for the dataset. The RBF kernel, in particular, is popular for its ability to capture complex, nonlinear relationships in the data. The process of selecting the optimal kernel function involves hyperparameter tuning, where the model is trained and evaluated using different kernels to identify the one that yields the highest classification accuracy.

Algorithm: SVM Algorithm

Step 1: Selecting an appropriate kernel function.
Step 2: Transforming the supplied data into a higher-dimensional space using the chosen kernel function.
Step 3: Optimizing the model parameters, including the regularization parameter C and kernel parameters, through techniques like grid search or random search.
Step 4: Training the SVM model on the transformed data to learn the decision boundary between classes.
Step 5: Evaluating the model's performance on a validation dataset using metrics such as accuracy, precision, recall, and F1-score.
Step 6: Fine-tuning the model parameters based on validation results to improve performance further.
In research, SVMs with kernel offer a powerful approach for early disease detection by effectively separating patients into disease and non-disease categories based on EHR data.

4. Experimental Results

4.1 Accuracy

Dataset	AI	GLMs	Proposed SVM
100	88.12	84.37	99.67
200	85.69	82.82	96.26
300	76.62	81.54	94.21
400	74.55	75.63	92.58
500	72.94	73.72	86.87

Table 1.Comparison Table of Accuracy

The Comparison table 1 of Accuracy Values explains the different values of existing algorithms (AI, GLMs) and proposed SVM. While comparing the Existing algorithm and proposed SVM, provides the better results. The existing algorithm values start from 72.94 to 88.12, 73.72 to 84.37 and proposed SVM values start from 86.87 to 99.67. The proposed gives the great results.

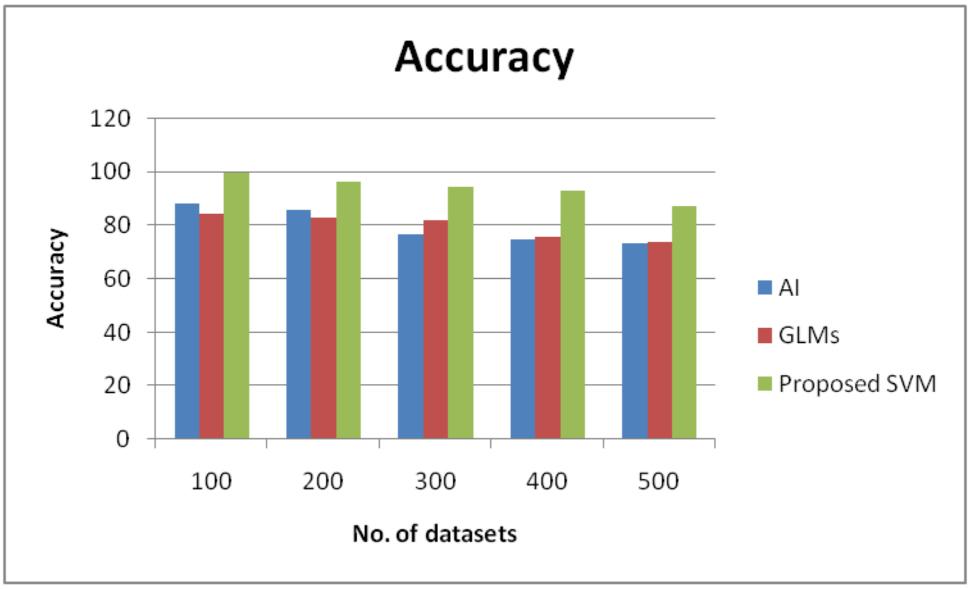

Figure 1.Comparison chart of Accuracy

The Figure 1 Shows the comparison chart of Accuracy demonstrates the existing1, existing 2 (AI, GLMs) and proposed SVM. The Y axis shows the accuracy in percentage, while the X axis indicates the number of datasets. The proposed SVM values outperform the current algorithm. The suggested SVM values start from 86.87 to 99.67, while the current algorithm values range from 72.94 to 88.12 and 73.72 to 84.37. The proposed produces excellent outcomes.

4.2 Recall

Dataset	AI	GLMs	Proposed SVM
100	0.73	0.81	0.84
200	0.74	0.77	0.91
300	0.81	0.74	0.95
400	0.85	0.73	0.96
500	0.86	0.72	0.98

Table 2.Comparison table of Recall

The Comparison table 2 of Recall Values explains the different values of existing algorithms (AI, GLMs) and proposed SVM. While comparing the Existing algorithm and proposed SVM, provides the better results. The existing algorithm values start from 0.73 to 0.86, 0.72 to 0.81 and proposed SVM values start from 0.84 to 0.98. The proposed gives the great results.

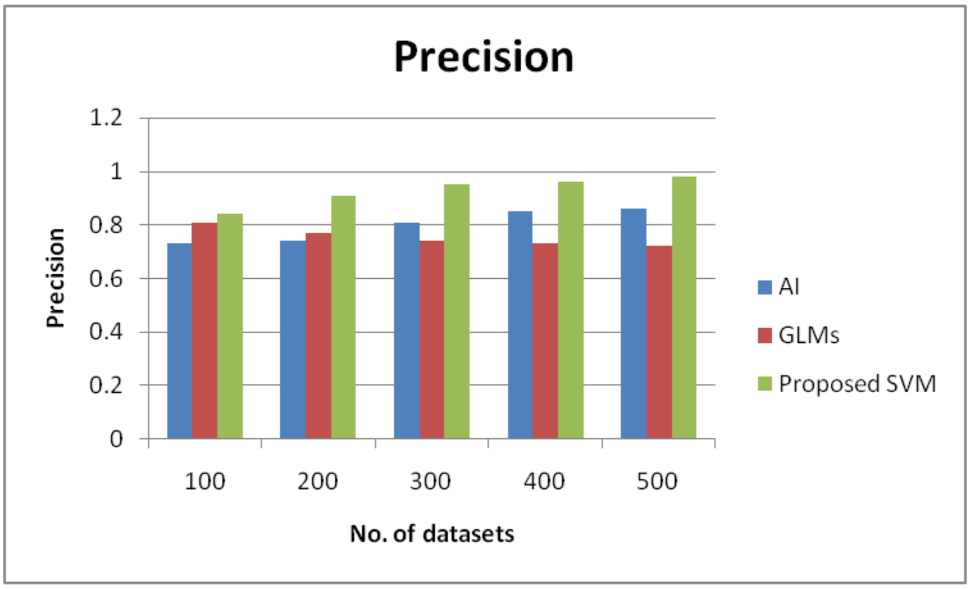

Figure 2.Comparison chart of Recall

The figure 2 shows recall comparison chart for the existing1, existing 2 (AI, GLMs), and suggested SVM. The Y axis shows the percentage of recall, and the X axis shows the number of datasets. The suggested SVM values outperform the current algorithm. The suggested SVM values begin at 0.84 and continue up to 0.98, while the current algorithm values range from 0.73 to 0.86 and 0.72 to 0.81. The suggested produces excellent outcomes.

5. Conclusion

In this paper, the integration of Support Vector Machines (SVM) into predictive modeling for early disease detection using electronic health records (EHRs) offers a structured approach to improving patient outcomes and reducing healthcare costs. By emphasizing data preprocessing, feature selection, and SVM model design, this methodology aims to effectively leverage EHR data for enhanced predictive accuracy. Through key steps such as data collection, preprocessing, and SVM model training with various kernel functions, this approach underscores the significance of leveraging sophisticated data mining methods in the medical field. Ultimately, this methodology holds promise for optimizing early disease detection capabilities and improving healthcare delivery.

References

1. Estiri H, Strasser ZH, Klann JG, Naseri P, Wagholikar KB, Murphy SN. Predicting COVID-19 mortality with electronic medical records. NPJ digital medicine. 2021 Feb 4;4(1):15.

2. Goh KH, Wang L, Yeow AY, Poh H, Li K, Yeow JJ, Tan GY. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nature communications. 2021 Jan 29;12(1):711.
3. Li Y, Rao S, Solares JR, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: transformer for electronic health records. Scientific reports. 2020 Apr 28;10(1):7155.
4. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, Lange J, Thiesson B. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature communications. 2020 Jul 31;11(1):3852.
5. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine. 2021 May 20;4(1):86.
6. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports. 2020 Jul 20;10(1):11981.
7. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA network open. 2020 Jan 3;3(1):e1918962-.
8. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, Strøm T, Chmura PJ, Heimann M, Dybdahl L, Spangsege L. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. The Lancet Digital Health. 2020 Apr 1;2(4):e179-91.
9. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association. 2020 Jul;27(7):1173-85.
10. Artzi NS, Shilo S, Hadar E, Rossman H, Barbash-Hazan S, Ben-Haroush A, Balicer RD, Feldman B, Wiznitzer A, Segal E. Prediction of gestational diabetes based on nationwide electronic health records. Nature medicine. 2020 Jan;26(1):71-6.