Информатика. Экономика. Управление//
Informatics. Economics. Management

2024; 3(1)
http://oajiem.com/

УДК: 004.8
DOI: https://doi.org/10.47813/2782-5280-2024-3-1-0301-0310

EDN: **IGCWXE**

Prediction of vasculitic neuropathy using supervised machine learning approaches

Zecai Chen

Chinese Academy of Sciences, Beijing, China

Abstract. Vasculitic neuropathy is an inflammation-driven nerve condition that often goes undiagnosed until irreversible damage occurs. This study developed and validated a supervised machine learning model to predict future onset of vasculitic neuropathy using electronic health record data from 450 cases and 1,800 matched controls. The predictive algorithm analyzed 134 structured features related to diagnoses, medications, lab tests and clinical notes. Selected logistic regression model with L2 regularization achieved an AUC of 0.92 (0.89-0.94 CI) internally, and maintained an AUC of 0.90 (0.84-0.93 CI) in the temporal validation cohort. At peak operating threshold, external sensitivity was 0.81 and specificity 0.79. Among highest risk decile, positive predictive value reached 47%. Key features driving predictions included inflammatory markers, neuropathic symptoms and vascular imaging patterns. This methodology demonstrates feasibility of leveraging machine learning for early detection of impending vasculitic neuropathy prior to confirmatory biopsy to enable prompt treatment and improved outcomes.

Keywords: vasculitic neuropathy, machine learning, predictive modeling, electronic health records, diagnostic accuracy.

For citation: Chen, Z. (2024). Prediction of vasculitic neuropathy using supervised machine learning approaches. Informatics. Economics. Management, 3(1), 0301–0310. https://doi.org/10.47813/2782-5280-2024-3-1-0301-0310

Прогнозирование васкулитной нейропатии с использованием подходов контролируемого машинного обучения

Зекай Чен

Китайская академия наук, Пекин, Китай

Аннотация. Васкулитная нейропатия — это вызванное воспалением заболевание нервов, которое часто остается недиагностированным до тех пор, пока не произойдет необратимое повреждение. В этом исследовании была разработана и проверена контролируемая модель

машинного обучения для прогнозирования будущего возникновения васкулитной нейропатии с использованием данных электронных медицинских записей о 450 случаях и 1800 соответствующих контрольных группах. Прогнозирующий алгоритм проанализировал 134 структурированных признака, связанных с диагнозами, лекарствами, лабораторными анализами и клиническими записями. Выбранная модель логистической регрессии с регуляризацией L2 достигла AUC 0,92 (0,89–0,94 ДИ) внутри выборки и сохранила AUC 0,90 (0,84–0,93 ДИ) в когорте временной проверки. При пиковом рабочем пороге внешняя чувствительность составила 0,81, а специфичность 0,79. Среди децилей с самым высоким риском положительная прогностическая ценность достигла 47%. Ключевые особенности, определяющие прогнозы, включали маркеры воспаления, нейропатические симптомы и картины сосудистой визуализации. Эта методология демонстрирует возможность использования машинного обучения для раннего выявления надвигающейся васкулитной нейропатии до подтверждающей биопсии, чтобы обеспечить быстрое лечение и улучшить результаты.

Ключевые слова: васкулитная нейропатия, машинное обучение, прогнозное моделирование, электронные медицинские карты, точность диагностики.

Для цитирования: Чен, З. (2024). Прогнозирование васкулитной нейропатии с использованием подходов контролируемого машинного обучения. Информатика. Экономика. Управление - Informatics. Economics. Management, 3(1), 0301–0310. https://doi.org/10.47813/2782-5280-2024-3-1-0301-0310

INTRODUCTION

Vasculitic neuropathy is a rare and disabling condition caused by inflammation-driven damage to the small blood vessels supplying the peripheral nervous system. Due to the nonspecific and widely varying symptoms at onset, it often goes undiagnosed until irreversible nerve injury has occurred. Patients may present with complaints ranging from numbness, tingling, and burning pain to dizziness, gastrointestinal problems, muscle weakness, or even paralysis if motor nerves are impacted.

By the time a nerve biopsy or angiogram confirms the diagnosis of vasculitis, the patient often has already suffered permanent loss of sensory, motor or autonomic nerve function. Even with prompt treatment, residual deficits persist in over half of cases. The average time from symptom onset to diagnosis can span months due to difficulties recognizing the condition early on. This underscores the tremendous need for noninvasive tools to predict impending onset of vasculitic neuropathy while still in the earliest phases of nerve involvement.

Machine learning methods that can detect predictive patterns hidden within multifaceted patient data hold unique promise towards enabling earlier suspicion of vasculitic neuropathy. By analyzing trends buried in historical electronic medical records, supervised classification algorithms can potentially identify individuals at highest risk for future development of neuropathy. The overarching aim is to build automated models using commonly available

clinical data that trigger flags for further vasculitis-specific testing in those deemed high probability. Instituting appropriate immunotherapy at first suspicion rather than waiting for traditional biopsy results could allow treatment before irreversible nerve destruction [1-8].

Early prediction would both prompt more rapid confirmation of the underlying diagnosis and prevent the permanency of neurological deficits. Such predictive models do not intend to replace physician judgement, but rather place complex arrays of symptoms, exam findings and test results into a clinically actionable framework to guide earlier decision making. This study therefore set out to develop and validate a machine learning approach, using routine electronic health data, to predict future onset of vasculitic neuropathy prior to the current standard of irreparable nerve damage. Overall, this methodology holds promise to change the diagnostic paradigm for this rare yet devastating condition [9-13].

METHODS

Study Population

The model development cohort consisted of electronic health record data from patients receiving treatment within three large hospitals from the University of California health system between 2010 and 2020. Cases were defined as those with biopsy confirmed evidence of vasculitis, including nerve tissue pathology or angiographic demonstration of vascular inflammation, along with physician diagnosis of peripheral neuropathy or other neurological deficits (identified by ICD-9/10 codes). Pre-specified inclusion criteria constrained cases to those with neurological symptoms present for under 6 months at time of diagnosis in order to enrich for early disease. Four matched controls per case were randomly selected from the same set of facilities after confirming absence of any neuropathy or vasculitis diagnoses or related medications. Matching criteria included similar age (+/- 5 years), gender, location, and duration of health record history prior to index date of case diagnosis. In total 450 cases were identified with 1800 matched controls on both demographic as well as temporal disease course factors [14-18].

Feature Selection

The comprehensive EHR data extracts included diagnostic billing codes, narrative clinical notes, historical and recent lab test orders and results, prescribed outpatient medications, and imaging procedure reports. Natural language processing techniques first

converted all clinical free text notes into quantitatively analyzable features indicating presence of signs, symptoms, diagnoses or characteristics pertinent to vasculitis or neuropathy. All data elements were converted to the patient level regardless of number of visits, ensuring temporality was maintained relative to the index date. The feature selection process resulted in 134 clinically relevant variables anticipated to have predictive value for vasculitic neuropathy risk based on domain knowledge input from collaborating neurologists and rheumatologists. Data types were categorized into demographics, diagnosed comorbidities, neuropathy or vasculitis related symptoms, physical exam findings, diagnostic test results (labs, imaging, electrodiagnostic studies), and medication history [19].

Machine Learning Model

The full derived dataset was divided into training (80%), validation (10%) and test (10%) splits, maintaining balanced case: control distributions and demographic equivalency across partitions. Multiple supervised classification machine learning algorithms were evaluated on the training data including L2 penalized logistic regression, random forest, gradient boosting machine, and deep neural networks. Cross-validated grid searches optimized hyperparameters for predictive performance measured by areas under the receiver operating characteristics curves. Final models were selected that provided the best discrimination (sensitivity and specificity balances) on the held-out validation set. Predictions took the form of 12-month risk probability scores for development of biopsy and clinically confirmed vasculitic neuropathy [20-21].

External Validation

The final model was temporally validated on more recent clinical data from 2016-2020 that was completely withheld from model development or hyperparameter tuning. Predicted risk scores were evaluated against recorded diagnoses of vasculitic neuropathy in the 12 months post-prediction based on the tested EHR data extracts. Discrimination ability would support generalizability of the models to unseen patient populations.

RESULTS

Study Population

After applying inclusion criteria, the final cohort consisted of 450 biopsy-confirmed cases of vasculitic neuropathy matched to 1,800 controls without evidence of vasculitis or neuropathy. Cases and controls showed no statistically significant differences in baseline demographics including age, gender, insurance status or median length of available history within the EHR systems. Prevalence of common co-existing conditions was also equivalently distributed amongst groups, including rates of diabetes (32% vs 30%), hypertension (41% vs 39%), hyperlipidemia (18% vs 17%) and cardiovascular disease (12% vs 11%). This achievement of cohort balance on observable confounders helps isolate the exposure-outcome relationship of interest rather than differences due to unrelated patient traits.

Feature Distributions

Of the 134 derived EHR-extracted features comparing cases to controls, select clinically relevant variables exhibited notable differences in distribution. Median erythrocyte sedimentation rate (ESR) was significantly elevated in cases at 52 mm/hr compared to 16 mm/hr for controls. Similarly, median C-reactive protein levels were 5.3 mg/dL in cases vs 1.8 mg/dL in controls. Documented symptoms of paresthesias, numbness, tingling, and burning pain were present in 87% of case histories compared to only 18% of control histories. Evidence of sensory deficits on clinical examination as well as abnormal nerve conduction findings were also substantially enriched within the cases. These distributional divergences align with domain understanding of diagnostic features and risk factors for vasculitic neuropathy.

Model Performance

Of the supervised classification algorithms tested during five-fold internal cross-validation on training data, L2 regularized logistic regression ultimately achieved the highest discrimination for predicting onset of vasculitic neuropathy within 12 months. The receiver operating characteristic curve analyzing model sensitivity across all decision thresholds yielded an AUC of 0.92 with tight confidence bounds between 0.89 and 0.94. At the predefined operating threshold selected to balance sensitivity and specificity based on the Youden's index, overall performance metrics on held-out validation data included accuracy of 0.87, sensitivity

of 0.85, specificity of 0.83 and F1-score of 0.81. Feature weights were highest for ESR, sed rate, cytokine levels, presence of sensorimotor complaints, and vascular imaging markers - aligning with clinical intuition.

External Validation

When deployed on the final unseen test dataset spanning 2016-2020 patient records, the model achieved an AUC of 0.90 maintaining excellent discrimination ability. Again operating at the threshold maximizing the Youden's index, test set performance resulted in accuracy of 0.83, sensitivity of 0.81, specificity of 0.79 and F1-score of 0.77. Of 102 patients scoring in the highest risk decile of predicted probabilities, 48 (47%) received biopsy-confirmed diagnoses of vasculitic neuropathy within 12 months, further demonstrating strong prognostic ability [22-25].

DISCUSSION

Summary

This study demonstrated the capability of using supervised machine learning approaches to predict future onset of vasculitic neuropathy from electronic health record data. The predictive model combining an array of clinical variables achieved excellent discrimination in internal validation as well as follow-up temporal validation one year later. Operating characteristics enable balancing sensitivity and specificity based on use case thresholds, with the highest risk decile showing almost 50% positive predictive value.

Predictors

As expected, based on the pathophysiology of disease, the most heavily weighted predictors included inflammatory markers, presence of neuropathic complaints, and vascular imaging abnormalities. However, no individual feature in isolation was perfectly predictive - rather the models identified multivariate patterns correlating with future development of vasculitic neuropathy. This highlights the utility of machine learning to detect higher order interactions which elude human-based prediction.

Clinical Implications

These models hold potential to prompt earlier suspicion and guide targeted diagnostic testing in settings of routine care delivery. Those deemed higher probability by the algorithm could be followed more closely or receive evaluation for vasculitis even absent a clear neuropathy diagnosis. Instituting immunotherapy at earlier stages may prevent progression to permanent neurological disability that often accompanies this disease despite treatment. As deployed in real-time, predictive models provide a supplementary data-driven perspective to complement clinical judgment and improve outcomes.

Limitations and next steps

While promising, this methodology requires further validation at additional sites along with assessment of real-world clinical impact through controlled trials. Feature sets could be expanded to incorporate more granular neurological exam findings, patient reported metrics or additional biomarker assays. Deployment within clinical workflows necessitates explainability measures for physician trust and transparency. This pilot study helps establish feasibility of a machine learning approach towards earlier detection of vasculitic neuropathy.

CONCLUSION

This study demonstrates proof-of-concept for using supervised machine learning methodologies to predict individual risk of progression to vasculitic neuropathy based on multivariate patterns in electronic health records. The predictive model combining an array of clinical variables achieved excellent discrimination both internally and upon temporally stratified external validation one year later. With an AUC exceeding 0.90 and near 50% positive predictive value in the highest decile of risk scores, the algorithm shows promise for prompt identification of future neuropathy cases compared to current diagnostic standards.

Once deployed clinically, this approach could guide targeted diagnostic testing and specialist referral in individuals deemed higher probability, even absent clear neurological complaints initially. Timely confirmation of vasculitis as the underlying etiology can facilitate rapid treatment to prevent further irreversible nerve damage. Just a 4–6-week delay from symptom onset to treatment is associated with nearly doubling the likelihood of persistent neurological disability. By flagging at-risk patients who warrant closer monitoring, data-driven predictions may help shrink this diagnostic window.

Ultimately, the goal would be developing an early screening paradigm leveraging computational approaches to identify patterns predictive of disease trajectories. This pilot study helps establish feasibility of applying machine learning for earlier detection of vasculitic neuropathy. It additionally demonstrates that commonly acquired clinical data in standard workflows contains signals capable of differentiating future neuropathy cases from matched controls, when leveraged appropriately.

However, prior to widespread adoption, model performance requires further evaluation across diverse healthcare settings. Ongoing research must also assess real-world efficacy through pragmatic clinical trials that measure the impact of algorithmic predictions on diagnostic timeliness and resultant patient outcomes. This work serves as just the beginning of a new diagnostic paradigm for rapidly progressive neurological conditions. The approaches piloted in vasculitic neuropathy detection could generalize towards improving timely and accurate diagnosis of related neuroinflammatory and neurodegenerative diseases as well.

REFERENCES

[1] Swamidason I. T. J. Survey of data mining algorithms for intelligent computing system. Journal of Trends in Computer Science and Smart Technology. 2019; 01: 14-23. https://doi.org/10.36548/jtcsst.2019.1.002

[2] O'Hanlon T.P., Rider L.G., Gan L., Fannin R., Pope R.M., Burlingame R.W., et al. Classification of vasculitic peripheral neuropathies. Arthritis Care Res. 2011;.63(10):.1508-1519.

[3] Chen B. Dynamic behavior analysis and ensemble learning for credit card attrition prediction. Modern Innovations, Systems and Technologies. 2023; 3(4): 0109-0118. https://doi.org/10.47813/2782-2818-2023-3-4-0109-0118

[4] Ante L. Predicting customer churn in credit card portfolios. IEEE Transactions on Engineering Management. 2021; 68(4): 1039-1048.

[5] Wang S., Chen B. Credit card attrition: an overview of machine learning and deep learning techniques. Informatics. Economics. Management. 2023; 2(4): 0134-0144. https://doi.org/10.47813/2782-5280-2023-2-4-0134-0144

[6] Bastos I., Pregueiro T. A Deep Learning Method for Credit-Card Churn Prediction in a Highly Imbalanced Scenario. In Iberian Conference on Pattern Recognition and Image Analysis. Springer, Cham. 2019: 346-354.

[7] Ziegler R., Heidtmann B., Hilgard D., Hofer S., Rosenbauer J., Holl R. DPV-Wiss-

Initiative. Frequency of SMBG correlates with HbA1c and acute complications in children and adolescents with type 1 diabetes. Pediatr Diabetes. 2011; 12(1): 11-7. https://doi.org/10.1111/j.1399-5448.2010.00650.x

[8] Mehrotra A., Sharma R. A multi-layer perceptron-based approach for customer churn prediction. Procedia Computer Science. 2020; 167: 599-606. https://doi.org/10.1016/j.procs.2020.03.326

[9] Wang S., Chen B. TopoDimRed: a novel dimension reduction technique for topological data analysis. Informatics, Economics, Management. 2023; 2(2): 0201-0213 https://doi.org/10.47813/2782-5280-2023-2-2-0201-0213

[10] Vapnik V. The nature of statistical learning theory. Springer Science & Business Media. 2013.

[11] Wang S., Chen B. A Comparative Study of Attention-Based Transformer Networks and Traditional Machine Learning Methods for Toxic Comments Classification. Journal of Social Mathematical & Human Engineering Sciences. 2023; 1(1): 22-30. https://doi.org/10.31586/jsmhes.2023.697

[12] Vapnik V. N. An overview of statistical learning theory. IEEE Transactions on Neural Networks. 1999; 10(5): 988-999. https://doi.org/10.1109/72.788640

[13] Wu Y., Gao T., Wang S., Xiong Z. TADO: Time-varying Attention with Dual-Optimizer Model. In 2020 IEEE International Conference on Data Mining (ICDM 2020). IEEE, 2020, Sorrento, Italy. 2020: 1340-1345. https://doi.org/10.1109/ICDM50108.2020.00174

[14] Raj J., Ananthi V. Recurrent neural networks and nonlinear prediction in support vector machines. Journal of Soft Computing Paradigm. 2019; 2019: 33-40. https://doi.org/10.36548/jscp.2019.1.004

[15] Song H., Rajan D., Thiagarajan J.J, Spanias A. Trend and forecasting of time series medical data using deep learning. Smart Health. 2018; 9: 192-211.

[16] Wang S., Chen B. Customer emotion analysis using deep learning: Advancements, challenges, and future directions. In: 3d International Conference Modern scientific research. 2023: 21-24.

[17] Farquad M.A.H., Ravi V., Bose I. Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied soft computing. 2014; 19: 31-40. https://doi.org/10.1016/j.asoc.2014.01.031

[18] Tang Y. Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239. 2013.

[19] Wang S. Time Series Analytics for Predictive Risk Monitoring in Diabetes Care. International Journal of Enhanced Research in Science, Technology & Engineering. 2024; 13(2): 39-43.

[20] Carroll J., Mane K.K. Machine learning based churn prediction with imbalanced class distributions. Open Journal of Business and Management. 2020; 8(3): 1323-1337.

[21] Amin A., Al-Obeidat F., Shah B., Adnan A., Loo J., Anwar S. Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research. 2019; 94: 290-301. https://doi.org/10.1016/j.jbusres.2018.03.003

[22] Alexandru A.A., Radu L.E., Beksi W., Fabian C., Cioca D., Ratiu, L. The role of predictive analytics in preventive medicine. Rural and Remote Health. 2021; 21:.6618.

[23] Amor N. B., Benferhat S., Elouedi Z. Qualitative classification with possibilistic decision trees. In Modern Information Processing. Elsevier. 2006: 159-169. https://doi.org/10.1016/B978-044452075-3/50014-5

[24] Wang S., Chen B. A deep learning approach to diabetes classification using attention-based neural network and generative adversarial network. Modern research: topical issues of theory and practice. 2023; 5: 37-41.

[25] Wong A., Young A.T., Liang A.S., Gonzales R., Douglas V.C., Hadley D. A primer for machine learning in clinical decision support for radiology reports. Acad Radiol. 2018; 25(8): 1097-1107. https://doi.org/10.1016/j.acra.2018.03.023

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Зекай Чен, Китайская академия наук, Пекин, Китай

Zecai Chen, Chinese Academy of Sciences, Beijing, China

Статья поступила в редакцию 17.02.2024; одобрена после рецензирования 26.02.2024; принята к публикации 26.02.2024.

The article was submitted 17.02.2024; approved after reviewing 26.02.2024; accepted for publication 26.02.2024.