ERSP: Prototype Parts Based Classification to Improve Interpretability of Deep Neural Network Predictive Model

April 24th, 2024

Trisha Gite

1 Research Context and Problem:

Various predictive models have been recently discovered that help predict motions in the human body via wearable sensor sleeves. However, these previous predictive models have downsides, causing them to have less accuracy in predicting motion. Such downsides include the lack of interpretability, privacy concerns, computational complexity, high cost, and more. Therefore, they have not addressed which predictive model is the most effective in predicting motions from a sensor sleeve. Accurate prediction of body motions and strokes by AI machine learning algorithms and predict models is crucial for timely medical intervention, potentially saving lives and minimizing long-term disabilities. Additionally, it optimizes resource allocation and enables personalized treatment strategies, leading to improved patient outcomes and cost savings in healthcare.

Therefore, our proposed solution uses Prototype parts based classification to propose some directions to enhance the Deep Neural Network's (DNN) accuracy of predicting human motion via sensor wearable slvees containing silver nanowire wires and stretchable electronics . Prototype Parts Based Classification is a special prototype layer where each unit of layer contains a detail of a decision the Deep Neural Network (DNN) can take. As the interpretability in DNN lacks the ability to see how the DNN makes its decisions, Prototype Parts Based Classification helps make the decision process more clear through the use of various prototypes for each possible decision. The process of how the DNN approaches the specific prototype decision will be further explained in detail. In the future of this research, this improvement of interpretability for the DNN predictive model can be transferred to read more data via the sensor wearable sleeves for predicting stroke conditions . This new data is still in the process of collection at our research lab.

Recent advancements in Artificial Intelligence (AI) have significantly enhanced the capabilities of predictive modeling and machine learning algorithms for stroke detection through wearable sensors. The research topic this proposal will be focusing on is creating a more effective predictive model that can read sensor data collected from the sensor sleeve and predict human motions at a high accuracy rate. It will be explaining one way to improve interpretability of the Deep Neural Network (DNN) which will potentially lead to a more accurate prediction of

human motions via the wearable sensor sleeves containing silver nanowire wires and stretchable electronics.

Li et al. (2018) introduced a neural network model combining an autoencoder with a prototype layer, where each unit stores a weight vector resembling encoded training inputs, originally applied in image classification [8]. Chen et al. (2019) built upon the work of Li et al. (2018) by introducing the prototypical part network (ProtoPNet), which includes a prototype layer storing prototypical parts derived from encoded training images. These parts, extracted as patches from convolutional neural network encodings, represent characteristic features for different image classes. When presented with an input image, the network evaluates its encoded form against the learned prototypical parts, generating a prototype activation map to identify both the location and degree of similarity with each part [9].

Related to predictive models, many studies also discuss how machine learning algorithms are being used in order to predict strokes in the human body at a high accuracy level. Yu et al.' (2023) study aimed to swiftly and accurately predict strokes in the elderly by utilizing real-time motion data, with a specific emphasis on early onset detection based on motion attribute values derived from walking [1]. Similarly, O'Brien et al., (2022) study aimed to test the value of high-resolution data from wireless, wearable motion sensors to predict post-stroke ambulation function following inpatient stroke rehabilitation, and found that models incorporating IMU data were more sensitive to patients who changed ambulation category, improving the recall of community ambulators at discharge from 85% to 89-93% [6].

Moving towards refining algorithms for precise stroke prediction, Lee et al. (2018) explored methods to make information presented via technology watches more closely related to the user's movements and strokes they might face, thereby enhancing the accuracy of stroke prediction algorithms [2].

Transitioning to studies that experiment with machine learning algorithms, Jalaja Jayalakshmi et al. (2021) tested various algorithms on a stroke prediction dataset, with J48 and adaBoost demonstrating high accuracy rates of 95.69% [4]. Similarly, Sharma et al., (2022) delves into the behaviors that cause strokes and tests five classification algorithms such as Naive Bayes Classification, Decision Tree, Random Forest, Multi Layer Perceptron, and JRip algorithms. After data mining the free available source "Stroke Prediction Data Set", they concluded that RandomForest had the highest accuracy (98.94%) in all three conditions of 70-30 percent ratio, ten-fold cross validation, and Feature Selection with Ranker Method using Information Gain Attribute Evaluator[5]. Thus contributing to the ongoing efforts in refining machine learning algorithms for stroke prediction and prevention.

Overall, these studies collectively underscore the evolving landscape of AI, predictive models, and wearable sensor technology in stroke and motion prediction. However, there are not many studies on finding and improving the existing predictive models such as the DNN to enhance its accuracy rate of predicting human motion via the sensor wearable sleeves. Therefore, we have come up with a new solution of integrating the *Prototype parts based classification* into

our research in order to improve the DNN's interpretability, leading to a higher accuracy rate of human motion prediction.

2 Current Research Status and Proposed Solution:

The main research "AI for Health" at North Carolina State University's overall aim as a whole is to predict strokes in the human body through the use of Artificial Intelligence. However, prior steps are needed to achieve this goal. One of these is the need to find a way on how to predict motions through the sensor data collected via the wearable sensor sleeves.

In our research, we are utilizing sensor sleeves equipped with advanced AgNW-based wearable technology. These sleeves incorporate highly conductive and stretchable conductors developed in PI Zhu's lab, boasting impressive conductivity exceeding 5,000 S cm^-1 and the ability to withstand strains of up to 50%. These sleeves feature a percolation network structure of AgNWs, allowing for high-resolution printing using techniques like screen printing and gravure printing. Within our study, we are deploying a range of sensors integrated into these sleeves, including strain/pressure sensors, electrophysiological electrodes, hydration sensors, temperature sensors, and antennas. These sensor sleeves have garnered significant interest from various sectors, including hospitals, industries, and the wider public, owing to their versatility and functionality. They come in two primary form-factors: adhesive patches for direct skin attachment and integrated fabrics, providing flexibility and ease of use in different applications. Notable examples within our research include highly stretchable strain sensors utilizing a dielectric layer for joint motion monitoring and AgNW-based dry electrodes offering superior signal quality without causing skin irritation or signal degradation, even during extended wear periods. These sensor sleeves represent a pivotal component of our investigation, enabling precise and reliable data collection for our research objectives [7].

The data collected through these sensor wearable sleeves is known as sensor data which will contain data in three categories: elbow movements, finger movements, and major muscles in the lower limbs. This data is still in the process of collection with an approximation finish date by the upcoming fall semester. Test subjects such as NC State students will be wearing the sensor wearable sleeves and will be making certain movements such as flexing arms, different leg movements, and various finger movements. The time the test subjects will be wearing this sensor serves for testing will be relatively short, approximately 5 minutes.

Along with this sensor data, we need a predictive model that will be able to analyze the data and correctly predict which motion the human is making. The predictive model we have decided to build on is called the Deep Neural Network (DNN). DNN is a type of artificial neural network (ANN) that consists of multiple layers between the input and output layers. These layers enable the network to learn complex patterns and representations from data. Deep neural networks have shown remarkable success in various tasks such as image recognition, natural language processing, and speech recognition [10]. They are characterized by their depth,

meaning they have many hidden layers, which allow them to learn hierarchical features from the input data. However, there exists few downsides to this model, including the lack of interpretability.

The "lack of interpretability" in deep neural networks (DNNs) refers to the challenge of understanding how these complex models make predictions. Unlike simpler models, DNNs operate as black boxes due to their high dimensionality, non-linear transformations, automated feature engineering, and intricate model architectures. This opacity makes it difficult to explain how specific inputs lead to particular outputs, hindering our ability to trust, verify, and debug these models, especially in critical applications. Addressing this challenge is crucial for ensuring transparency, accountability, and user trust in DNN-based systems, motivating ongoing research into methods for improving interpretability, such as visualization techniques, model-agnostic interpretation methods, and interpretable model architectures.

Therefore, as our proposed solution we have decided to come up with a way to improve the DNN's lack of interpretability. The proposed solution to help the problem of interpretability in DNNs is through the use of *Prototype Parts Based Classification.* Prototype-based classification is a machine learning approach where instances are represented as prototypes, which are typically a subset of the training data. In this context, "parts-based classification" suggests that each prototype represents a particular aspect or part of the data, rather than the entire data distribution. Each prototype serves as a representative example of a class or category. When classifying a new instance, its similarity to the prototypes is evaluated, and it is assigned to the class associated with the most similar prototype. This approach is particularly useful when dealing with high-dimensional data or complex data distributions, as it can help simplify the classification task by focusing on essential aspects or parts of the data [10] . In summation, it helps humans to understand how the DNN reached its conclusion to correctly predict which motion the human is making after reading in the sensor data.

Later, this proposed solution of implementing Prototype Parts Based Classification will be used to analyze and classify more data and has the potential to help predict strokes in the human body, the ultimate aim of this research "AI for Health" at North Carolina State University.

3. Evaluation Plan and Implementation Plan

In this section, how Prototype Parts Based Classification will be integrated specifically into our research will be explained with great detail. As explained earlier, our sensor data is sorted into three primary categories: elbow movements, finger movements, and major muscles in the lower limbs. To facilitate effective analysis, the incoming data undergoes extensive mining and quality/control checks to ensure its integrity and accuracy within the controlled environment of our research lab.

The main integration aspect of our approach lies in the actual employment of Prototype Parts Based Classification as a sophisticated methodology. It incorporates a specialized prototype layer within our Deep Neural Network (DNN). This distinctive layer contains individual units, or prototypes, each meticulously tailored to encapsulate the characteristic ranges associated with the aforementioned categories of sensor data. These prototypes serve as identifying templates, tracing various human motions with remarkable precision.

By the use of this classification mechanism, our system can observe a spectrum of human movements with great potential for accuracy. For instance, it can identify actions such as walking (characterized by minimal finger movement, swaying of the elbow, and a consistent limb motion pattern), sitting (marked by the absence of movement across all categories), and standing (where the limb muscles do not move). Additionally, it can also identify subtler motions like unstable limb movement (manifested as trembling in the lower limbs), open hand gestures (with fingers extended), closed tight gestures (where all fingers are tightly clenched), and trembling elbow movements (exhibited as trembling in the elbow) which will play an important role when predicting strokes in the human body.

To evaluate the effectiveness of the Prototype Parts Based Classification integrated into the Deep Neural Network (DNN) for predicting human motion accurately and improving interpretability, several evaluation measures will be taking place. Baseline models, including traditional machine learning algorithms or the DNN without the prototype layer, will serve as comparison points to test the performance enhancement brought by Prototype Parts Based Classification. DNNs have a predictive accuracy rate of about 91-97% for predicting human motions via sensor wearable sleeves without the integration of the Prototype Parts Based Classification [11]. Accuracy metrics will measure the model's ability to correctly identify various types of human motions explained above, while interpretability will be assessed qualitatively and quantitatively through visualization techniques and user studies. Through these measures, we aim to measure the Prototype Parts Based Classification efficiency in enhancing both prediction accuracy and interpretability in human motion prediction from wearable sensor data.

The root of this approach lies in the alignment of the ranges extracted from the sensor data with the corresponding prototypes. Through careful comparison, the system determines the most suitable prototype, thus determining the specific type of human motion being exhibited. This integration of Prototype Parts Based Classification into our DNN augments the

decision-making process, enhancing the interpretability, efficiency, and an increased accuracy prediction rate of the overall system. By equipping our DNN with this interpretative capability, we not only enhance the accuracy of human motion prediction but also gain valuable insights into the underlying mechanisms governing the decision-making process. It provides transparency into how the DNN arrives at its conclusions, thereby boosting confidence in the reliability and efficacy of our research findings therefore increasing the interpretability. In conclusion, the incorporation of Prototype Parts Based Classification represents a pivotal advancement in our quest to predict human motion accurately and effectively from sensor data while increasing the interpretability.

Evaluation Timeline

- Fall Semester 2024:

    - Week 1:
        * clean and mine the received fresh data
    - Week 2:
        * Conduct quality/control checks on the collected sensor data with research team
        * Begin integration of Prototype Parts Based Classification into the DNN
    - Week 3:
        * Continue integration process of Prototype Parts Based Classification
        * Conduct initial tests to ensure proper functioning of the integrated system
    - Week 4:
        * Refine Prototype Parts Based Classification based on initial test results
        * Collect feedback from research team for further improvements
    - Week 5:
        * Continue refining Prototype Parts Based Classification
        * Conduct additional tests to validate accuracy and interpretability improvements
    - Week 6:
        * Analyze test results and make necessary adjustments to the system
    - Week 7:
        * Conduct thorough evaluation of Prototype Parts Based Classification performance
        * Address any remaining issues or challenges
    - Week 8:
        * Finalize integration of Prototype Parts Based Classification into the DNN
        * Conduct tests on the prototypes being concluded by the DNN, testing accuracy
    - Week 9:
        * Prepare for presentation of preliminary results to research team
    - Week 10:
        * Begin drafting research paper outlining methodology and results
    - Week 11:
        * Complete Conclusion section
    - Week 12:
        * Create poster
    - Week 13:
        * Final revisions to paper, get final feedback from Research Mentor
    - Week 14:
        * Present
        * Plan for future research and potential applications of integrated system

References

[1] Yu, J., Park, S., Ho, C. M. B., Kwon, S.-H., Cho, K.-H., and Lee, Y. S. Aibased stroke prediction system using body motion biosignals during walking. The Journal of Supercomputing (2022), 1–23

[2] Lee, S. I., Adans-Dester, C. P., Grimaldi, M., Dowling, A. V., Horak, P. C., Black-Schaffer, R.M., Bonato, P., and Gwin, J. T. Enabling stroke rehabilitation in home and community settings: a wearable sensor based approach for upper-limb motor training. IEEE journal of translational engineering in health and medicine 6 (2018), 1–11

[3] Sett, N., Mac Namee, B., Calvo, F., Caulfield, B., Costello, J., Donnelly,S. C., Dorn, J. F., Jeay, L., Keogh, A., McManus, K., et al. Are you in pain? predicting pain and stiffness from wearable sensor activity data. 183–197

[4] JalajaJayalakshmi, V., Geetha, V., and Ijaz, M. M. Analysis and prediction of stroke using machine learning algorithms. 1–5

[5] Sharma, C., Sharma, S., Kumar, M., and Sodhi, A. Early stroke prediction using machine learning. 890–894

[6] O'Brien, M. K., Shin, S. Y., Khazanchi, R., Fanton, M., Lieber, R. L., Ghaffari, R., Rogers, J. A., and Jayaraman, A. Wearable sensors improve prediction of post-stroke walking function following inpatient rehabilitation. IEEE Journal of Translational Engineering in Health and Medicine 10 (2022), 1–11

[7] Patel, S., Park, H., Bonato, P., Chan, L., and Rodgers, M., 2012, "A review of wearable sensors and systems with application in rehabilitation," J. Neuroeng. Rehabil., **9**(1), p. 21.

[8] Bing Liu, Wynne Hsu, Yiming Ma, et al. Integrating classification and association rule mining. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), volume 98, pages 80–86, 1998.

[9] Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Proceedings of Neural Information Processing Systems (NeurIPS), volume 31, 2018.

[10] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys, 16, 1-85.

[11] Song, Y., Taylor, W., Ge, Y., Usman, M., Imran, M. A., and Abbasi, Q. H. Evaluation of deep learning models in contactless human motion detection system for next generation healthcare. Scientific Reports 12, 1 (2022), 21592