Predicting Freezing of Gait from Fall Reports

Masoud Nateghi
Emory University
Department of Biomedical Informatics
masoud.nateghi@emory.edu

Abstract

Freezing of Gait (FoG) remains one of the poorly-understood symptoms of Parkinson's Disease (PD) that requires further investigation. This symptom significantly contributes to falls among elderly patients. Consequently, identifying the key factors triggering FoG can be immensely valuable in alerting patients to potential hazards.

This study aims to simplify the challenging task of determining whether FoG is the cause of a fall, employing Natural Language Processing (NLP) techniques. Our experiments involved various machine learning models applied to fall self-report dataset. We discovered that the utilization of word embedding models, such as Word2Vec and gLoVe, combined with tf-idf vectorization on n-grams, can provide valuable insights for training an automatic classification model.

The most promising results were obtained using an SVM model, achieving an 84% accuracy rate and a macro-averaged F1 score of 84% on the test dataset.

1 Introduction

Freezing of Gait (FoG) is a common symptom in Parkinson's Disease (PD) about which we have limited understanding (Moore et al., 2007; Giladi et al., 1997). FoG can be defined as "brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk" (4th International Workshop on Freezing of Gait). During such episodes, the patient's upper body momentum continues, while the patient remains unable to step forward, often leading to a fall. Risk factors for FoG include sex, left-sided disease, early gait abnormalities, a higher daily dose of levodopa, and other nonmotor symptoms such as hallucinations, depression, and anxiety (Rahimpour et al., 2021). The duration of each FoG episode can vary from a few seconds to more than 30 seconds (Schaafsma et al., 2003), and these falls are more likely to occur while walking, turning, passing through narrow passages, or approaching destinations (Nutt et al., 2011). FoG can result in falls, diminish the quality of life, and render the individual dependent on other family members for daily routine tasks (Kerr et al., 2010).

Individuals with PD are at a higher risk of falling compared to elderly individuals without the disease (Bloem et al., 2001; Stack and Ashburn, 1999). Therefore, it is crucial to accurately document the circumstances surrounding FoG occurrences, as risk factors may differ between these two groups (Powell et al., 2023). Capturing detailed contextual information in text format is essential for identifying the root cause of falls in patients (Stack and Ashburn, 1999). Natural Language Processing (NLP) techniques have proven to be invaluable for extracting biomedical information from unstructured health records such as electronic health records (EHR) (Houssein et al., 2021). In this paper, our focus is on developing an automated classifier that utilizes textual fall descriptions to predict whether the cause of the fall is FoG.

2 Methods

2.1 Dataset

In a recent study (McKay et al., 2021), the target population with PD was carefully observed over the course of a year. All participants underwent a comprehensive cognitive and motor assessment to identify potential fall risks during a single study visit. After enrollment, the study participants were prospectively monitored for any falls over a period of one year. Throughout the observation period, participants were regularly asked about any falls and their circumstances at monthly intervals via mail, email, phone, or text, as preferred by the participant. Participants recorded the details of any subsequent falls, which were then verified by the study staff. Approximately one-third of the enrolled participants experienced a fall during the observation period.

Each record was accompanied by PD risk factors such as age, sex, duration of PD, and so forth, along with a detailed description of the circumstances surrounding the fall. These falls were classified by an expert as either FoG or a normal fall. The training dataset consisted of a total of 299 patients, while the test dataset comprised 71 randomly selected patients.

2.2 Data Preprocessing

In this study, we focused exclusively on the fall description column of the dataset, using it as the plain text input for our classifier, while disregarding all other potential risk factors. Following this, we proceeded to eliminate all participants who had not provided a description of their fall in the training (3 samples) and test (2 samples) datasets, as indicated by the "N/A" entries within the descriptions.

Furthermore, as part of the text preprocessing stage, we implemented several key procedures. First, we converted each sentence to lowercase and applied stemming to individual words using the Porter stemmer. Notably, we chose to remove only periods and commas as punctuation, without removing any stop words. Subsequently, we employed the generation of n-grams (n = 1, 2, 3) and CMU (Carnegie Mellon University) word clusters(Owoputi et al., 2013) on pre-processed sentences. Leveraging the CMU word clusters, our methodology involved matching each word in the sentence with the vocabulary words in cluster and their respective cluster IDs, thereby replacing the original words with their corresponding cluster IDs. It is crucial to note that we excluded any words that did not appear in the CMU word clusters.

Having transformed all words into their respective cluster IDs, we treated the new string in the same manner as the original string, enabling us to generate n-grams and build a vocabulary for subsequent processing, such as vectorization.

Regarding vectorization, we adopted three distinct approaches. First, the straightforward and convenient technique of count vectorization involved counting the occurrences of each word in the vocabulary within the given sentence. This method resulted in the creation of an $n \times p$ feature matrix, where n represents the number of samples in the dataset and p denotes the number of words in the vocabulary.

Additionally, we employed the term frequency-inverse document frequency (tf-idf) method, where the "tf" component is identical to the count vectorizer, while the "idf" component was determined by the total number of sentences in which the vocabulary word appeared. By dividing the term frequency by the document frequency, we could obtain the tf-idf value. To create our features using these vectorizers, we set the maximum features at 10,000.

Lastly, we incorporated pre-trained models to extract word embeddings, similar to the concept of CMU word clusters, where words exhibiting similarity possessed closely related embeddings. We used word2vec and GloVe word embedding models with a feature size of 300. For each sentence, we computed the word embedding of each individual word and then computed the average word embeddings across the entire sentence.

Overall, our feature extraction process resulted in the acquisition of 4,000 features from n-grams and CMU word clusters through count vector and tf-idf vectorizers, alongside an additional 600 features obtained from the word embedding models which were all normalized in the last stage.

2.3 classification

Given that our task involves predicting whether a fall is a result of Freezing of Gait (FoG) or a normal fall, we approached the problem as a binary classification task, relying solely on the text descriptions. Our analysis encompassed a range of binary classifiers, including a Gaussian Naïve Bayes classifier as a baseline, Support Vector Machines (SVM), XGBoost, Random Forests, Logistic Regression, K-Nearest Neighbours (KNN), and an ensemble consisting of SVM, Random Forest, and KNN, utilizing a hard voting procedure.

Notably, considering the utilization of 6 feature sets, we encountered a total of 63 different combinations for feature selection. To identify the most optimal feature sets, we employed SVM, Logistic Regression, and KNN classifiers, individually evaluating each combination of feature sets through 5-fold cross-validation (CV) to fine-tune the corresponding hyperparameters. Through rigorous experimentation and iterative refinements, we determined that the highest accuracy was achieved when employing feature sets generated by tf-idf on n-grams, along with word embeddings from Word2Vec and GloVe.

Due to the extensive time consumption associated with repeating the same process for XGBoost and Random Forests, we opted to utilize the identified optimal feature sets for these models, subsequently fine-tuning their hyperparameters using 5-fold CV. Finally, for the ensemble model, we employed the previously identified optimal parameters, having determined them earlier for SVM, Random Forest, and KNN.

3 Results

3.1 Performance of Classifiers

Each model was individually trained on the dataset consisting of 296 patients in the training set. These models were subsequently evaluated using 5-fold cross-validation. The classifiers were then tested with the optimal feature set and hyperparameters on the separate test dataset, which included 69 patients.

To compare the performance of the various models, we considered metrics such as accuracy, micro-averaged F1 scores for both positive (FoG) and negative (nFoG) classes, as well as macro-averaged F1 scores which are all denoted in Table 1. Notably, the SVM model outperformed all other classifiers in terms of performance. As illustrated in Figure 1, we observe an overall improvement in performance with an increase in the amount of data used. For the optimal SVM model, we find that performance reaches a saturation point when approximately 70% of the training data is utilized, equivalent to 215 patients. Additionally, laborious and time-consuming task of manually annotating more data leads to only a slight enhancement in the model's performance.

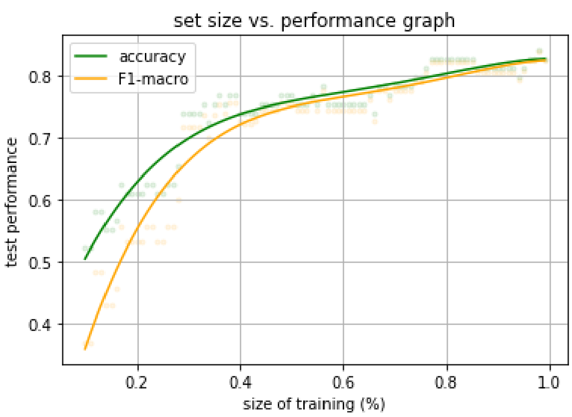

Figure 1: Interpolation of the performance of the SVM classifier for accuracy and micro-averaged F1 score, employing a subset of the complete training dataset.

3.2 Ablation Study

In this stage, we conducted an ablation study on SVM classifier, given its superior performance compared to other classifiers. We systematically removed one feature at a time and monitored the changes in the metrics used in Table 1. We noticed a decrease in all metrics when each feature was removed, except for the glove word embeddings, for which no decrease was observed (Table 2).

4 Discussion

Our study was centered on the utilization of diverse machine learning models to address the task of binary classification. Specifically, we aimed to determine whether a fall stemmed from freezing of gait (FoG) or was a normal fall within the elderly population, utilizing fall descriptions presented in plain text format. Our findings underscored the notable efficacy of the SVM model in this specific task, surpassing other classifiers with an impressive accuracy of 84% and a macro-averaged F1 score of 84%.

Significantly, we uncovered that using 70% of the training dataset (215 patients) could produce satisfactory results. Our experiments with feature sets highlighted the superiority of high-level feature extraction methods like word embeddings, such as Word2Vec and gLoVe, compared to lower-level approaches like count vectors and tf-idf for achieving optimal performance.

Moreover, our research emphasizes the critical role of leveraging unstructured health records to extract crucial biomedical insights, underscoring the substantial implications of natural language processing (NLP) in the biomedical domain. NLP stands as a valuable tool, enabling the extraction of pertinent information from textual data and facilitating accurate differentiation for predicting the nature of the fall.

5 Conclusion

This study highlights the potential of employing machine learning models to predict the type of falls in the elderly population with Parkinson's Disease (PD). It was observed that advanced features like word embeddings proved highly effective in extracting informative features from unstructured textual data, surpassing the performance of conventional vectorizing methods such as count vectorizer and tf-idf. Utilizing these features, it became evident

Model	Accuracy	F1-micro FoG	F1-micro nFoG	F1-macro
Gaussian Naïve Bayes	0.56	0.56	0.62	0.59
XGBoost	0.70	0.66	0.73	0.69
Logistic Regression	0.74	0.74	0.74	0.74
KNN	0.78	0.78	0.79	0.78
Random Forest	0.80	0.77	0.82	0.79
ensemble	0.83	0.80	0.85	0.82
SVM	**0.84**	**0.82**	**0.86**	**0.84**

Table 1: Performance of various classifiers on the fall dataset. 'FoG' denotes the positive class, while 'nFoG' represents the negative class.

Removed Feature	Accuracy	F1-micro FoG	F1-micro nFoG	F1-macro
tf-idf + n-grams	0.80	0.78	0.81	0.80
Word2Vec Embeddings	0.81	0.79	0.83	0.81
GloVe Embeddings	**0.84**	**0.82**	**0.86**	**0.84**

Table 2: Performance of SVM model on ablation study.

that the SVM model yielded promising outcomes in this context, outperforming other classifiers.

6 Supplementary

All the source codes along with information for models and tuning their hyperparameters can be found at: BMI550/Assignment2.

References

4th International Workshop on Freezing of Gait.

B. R. Bloem, Y. A. Grimbergen, M. Cramer, M. Willemsen, and A. H. Zwinderman. 2001. Prospective assessment of falls in Parkinson's disease. *J Neurol*, 248(11):950–958.

N. Giladi, R. Kao, and S. Fahn. 1997. Freezing phenomenon in patients with parkinsonian syndromes. *Mov Disord*, 12(3):302–305.

Essam H. Houssein, Rehab E. Mohamed, and Abdelmgeid A. Ali. 2021. Machine learning techniques for biomedical natural language processing: A comprehensive review. *IEEE Access*, 9:140628–140653.

G. K. Kerr, C. J. Worringham, M. H. Cole, P. F. Lacherez, J. M. Wood, and P. A. Silburn. 2010. Predictors of future falls in Parkinson disease. *Neurology*, 75(2):116–124.

J. L. McKay, K. C. Lang, S. M. Bong, M. E. Hackney, S. A. Factor, and L. H. Ting. 2021. Abnormal center of mass feedback responses during balance: A potential biomarker of falls in Parkinson's disease. *PLoS One*, 16(5):e0252119.

O. Moore, C. Peretz, and N. Giladi. 2007. Freezing of gait affects quality of life of peoples with Parkinson's disease beyond its relationships with mobility and gait. *Mov Disord*, 22(15):2192–2195.

J. G. Nutt, B. R. Bloem, N. Giladi, M. Hallett, F. B. Horak, and A. Nieuwboer. 2011. Freezing of gait: moving forward on a mysterious clinical phenomenon. *Lancet Neurol*, 10(8):734–744.

O. Owoputi, B. O'Connor, C. Dyer, Kevin Gimpel, N. Schneider, and N.A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. *Proceedings of NAACL-HLT*, 2013:380–390.

Jeanne M. Powell, Yuting Guo, Abeed Sarker, and J. Lucas McKay. 2023. Classification of fall types in parkinson's disease from self-report data using natural language processing. In *Artificial Intelligence in Medicine*, pages 163–172, Cham. Springer Nature Switzerland.

S. Rahimpour, W. Gaztanaga, A. P. Yadav, S. J. Chang, M. O. Krucoff, I. Cajigas, D. A. Turner, and D. D. Wang. 2021. Freezing of Gait in Parkinson's Disease: Invasive and Noninvasive Neuromodulation. *Neuromodulation*, 24(5):829–842.

J. D. Schaafsma, Y. Balash, T. Gurevich, A. L. Bartels, J. M. Hausdorff, and N. Giladi. 2003. Characterization of freezing of gait subtypes and the response of each to levodopa in Parkinson's disease. *Eur J Neurol*, 10(4):391–398.

E. Stack and A. Ashburn. 1999. Fall events described by people with Parkinson's disease: implications for clinical interviewing and the research agenda. *Physiother Res Int*, 4(3):190–200.