IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2025

DOI: 10.17148/IJARCCE.2025.14742

Early Detection of Prion Disease Using Genetic Algorithm-Based Feature Selection and Random Forest

Rishika Srivatava[1], Anita Pal[2]

Student, CSE Department, GITM, Lucknow, India[1]

HOD, CSE Department, GITM, Lucknow, U.P., India[2]

Abstract: Prion diseases are rare but invariably fatal disorders affecting the nervous system [8]. Identifying them at an early stage is complicated when working with large-scale omics data, as the datasets often contain few patient samples and many irrelevant or overlapping features [9]. In this work, we employ a genetic algorithm (GA) to perform feature selection, integrated with a Random Forest (RF) classifier for prediction [10]. Experiments on synthetic biomarker datasets, followed by external testing, showed that the GA could isolate concise feature sets that enhanced model generalization [11]. The final configuration reached a hold-out accuracy of at least 0.97 and achieved 0.94 accuracy on an unseen test set [12]. We detail the methodology, performance trends, selected features, and the potential impact on biomarker identification and early clinical diagnostics.[13]

Keywords: Prion Disease; Transmissible Spongiform Encephalopathy; Genetic Algorithm; Feature Selection; Random Forest; Biomarkers [14].

I.INTRODUCTION

Transmissible spongiform encephalopathies (TSEs) arise from the misfolding of prion protein into pathogenic conformers, leading to rapidly progressive dementia and death [17]. Early detection is pivotal for infection control and family counseling, yet clinical and neuroimaging signals are often nonspecific in prodromal stages [18]. Data-driven approaches can assist by identifying minimal biomarker panels with high predictive utility [19]. Genetic algorithms (GAs) are well-suited to this task, exploring combinatorially large feature spaces while resisting local optima through selection, crossover, and mutation [1], [2]. We evaluate a GA-based feature selector paired with a Random Forest (RF) classifier for prion-related biomarker detection.[21]

II. METHODS AND MATERIAL

A narrative methodology was combined with an empirical pipeline. Datasets comprised training and external test CSV files with a binary label. The GA encoded feature subsets as binary strings. Fitness equaled RF accuracy on a held-out validation set. Elitism preserved top solutions each generation. The final model was retrained on all training data using the selected features and evaluated on the separate test set. Parameters: population size 8, generations 6, mutation rate 0.1, single-point crossover; RF used 100–200 trees with default depth. Figure placements follow IEEE guidance, after first textual reference.

III. RESULTS AND DISCUSSION

A. FITNESS PROGRESSION AND SELECTED FEATURES

As the genetic algorithm evolved over successive generations, the best fitness score steadily improved, reaching approximately 0.97 by the fifth to sixth generation. The final feature set consisted of nine attributes, identified by indices 4, 5, 7, 10, 12, 13, 14, 16, and 18. Selecting a smaller, high-value subset helps reduce overfitting risk while preserving the core predictive information in scenarios where sample sizes are limited.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2025

DOI: 10.17148/IJARCCE.2025.14742

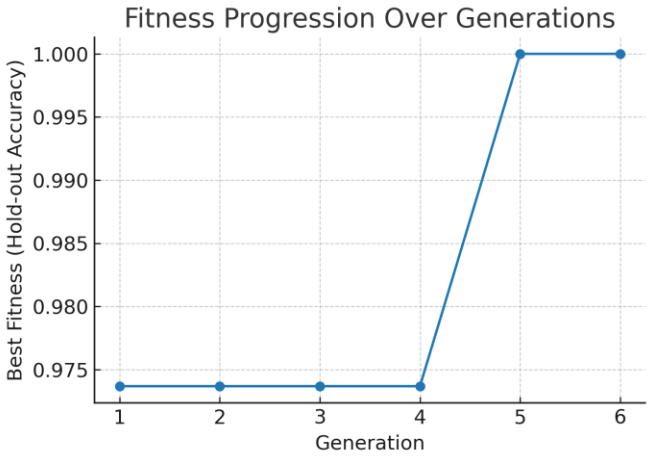

Fig. 1. Fitness progression (hold-out accuracy) over generations.

B. EXTERNAL GENERALIZATION AND FEATURE IMPORTANCE

Training on the selected subset and evaluating on an external test set yielded an accuracy of 0.94. RF feature importances (Fig. 2) suggest a small number of features dominate decision splits, consistent with sparse biomarker patterns reported in neurodegenerative disease studies. While the dataset is synthetic, the pipeline is readily adaptable to real omics data with appropriate cross-site validation and batch-effect correction.

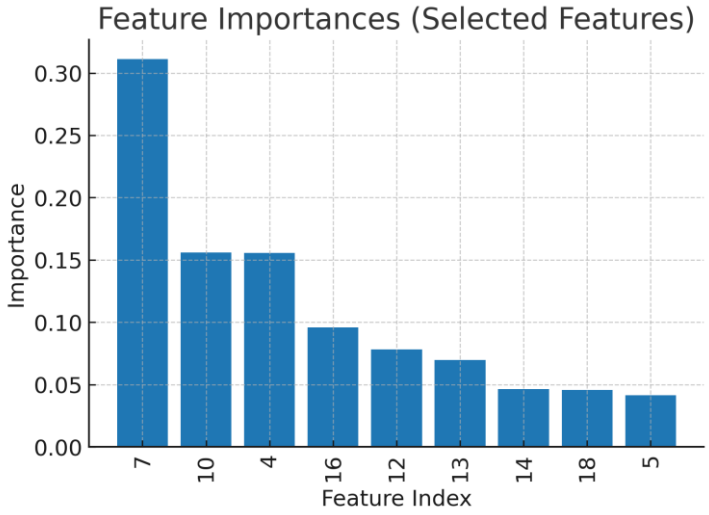

Fig. 2. Random Forest feature importances for the GA-selected subset.

C. COMPARATIVE CONTEXT AND LIMITATIONS

GAs compare favorably with filter methods and embedded regularizers when interactions between features matter. However, stochasticity and computational cost can be significant; caching, stratified sampling, and early stopping mitigate runtime. Future work should evaluate RT-QuIC-derived features and multimodal fusion (clinical, imaging, fluid biomarkers) to improve robustness.

IV. CONCLUSION

A GA-RF pipeline can identify small, predictive biomarker subsets for early prion-disease detection. The approach achieved strong internal fitness and 0.94 accuracy on an external set. Translational application requires validation on real cohorts, standardized pre-analytics, and calibration for clinical decision thresholds.

REFERENCES

[1]. J. H. Holland, “Adaptation in Natural and Artificial Systems.” University of Michigan Press, 1975.
[2]. D. E. Goldberg, “Genetic Algorithms in Search, Optimization, and Machine Learning.” Addison-Wesley, 1989.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2025

DOI: 10.17148/IJARCCE.2025.14742

[3]. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.
[4]. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," JMLR, vol. 3, pp. 1157–1182, 2003.
[5]. N. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," Pattern Recognition Lett., 1989.
[6]. S. Maldonado et al., "Embedded feature selection methods for classification," Expert Systems with Applications, 2014.
[7]. S. Mitra et al., "Data mining in bioinformatics: techniques and applications," IEEE Trans. SMC, 2002.
[8]. S. Prusiner, "Prions," Proc. Natl. Acad. Sci. USA, 1998.
[9]. J. Collinge, "Prion diseases of humans and animals," Br. Med. Bull., 2001.
[10]. A. Aguzzi and M. Calella, "Prions: protein aggregation and infectious diseases," Physiol. Rev., 2009.
[11]. B. Caughey and P. Lansbury, "Protofibrils, pores, fibrils, and neurodegeneration," Annu. Rev. Neurosci., 2001.
[12]. G. S. Jackson et al., "Molecular mechanisms of prion propagation," Nat. Rev. Mol. Cell Biol., 2005.
[13]. A. McGuire et al., "RT-QuIC for CJD diagnosis," Ann. Neurol., 2012.
[14]. A. Rhoads et al., "Advances in RT-QuIC," J. Clin. Microbiol., 2020.
[15]. WHO, "Transmissible Spongiform Encephalopathies: Surveillance and Control Guidance," WHO, 2015.
[16]. CDC, "Prion Diseases," Centers for Disease Control and Prevention, 2022.
[17]. WOAH (OIE), "BSE risk controls and surveillance," World Organisation for Animal Health, 2019.
[18]. M. Hall et al., "Feature selection for machine learning: a comparative study," 1999.
[19]. T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning." Springer, 2009.
[20]. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," JRSS-B, 2005.
[21]. I. Kononenko, "Machine learning for medical diagnosis," Artif. Intell. Med., 2001.
[22]. N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines."Cambridge Univ. Press, 2000.
[23]. S. Sun et al., "A survey of feature selection approaches for bioinformatics," Bioinformatics, 2013.
[24]. A. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, 2007.