Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2025) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Genetic Brain Disease Classification using Machine Learning, Deep Learning & Custom Learning Models

Aryan Kothari[a], VRNS Nikhil[a], Namana Rohit[a], Vasavi C.S.[b], Karthikeyan B[c]

[a]*Department of Computer Science and Engineering, Amrita School of Computing, Bengaluru, Amrita Vishwa Vidhyapeetham, India*
[b]*Department of Artificial Intelligence, Amrita School of Artificial Intelligence, Bengaluru, Amrita Vishwa Vidyapeetham, India*
[c]*Department of Embedded Technology, School of Electronics Engineering, Vellore Institute of Technology, Vellore, India*

Abstract

This research brings out a new approach to identifying genetic brain diseases in their early stages by employing DNA sequence embeddings. This dataset has employed the largest dataset to date(5000 sequences). Our method maps raw DNA sequences into a structured vector representation, which includes important structural and functional genomic features. These embeddings are used as input to more sophisticated classifiers such as the proposed Stacking Classifier, CNNs etc, to learn patterns that may point towards different brain diseases. From the results of our experiments, we can conclude that the proposed Stacking Classifier, trained according to the particularities of a specific domain, has an accuracy of 94%, which is nearly similar to the CNN model's performance. This equivalence in performance establishes the effectiveness of the proposed ensemble system design that incorporates multiple base learners coherently and synergistically.

Keywords: Brain Diseases; DNA sequence; DNABert Embeddings; deep learning; Machine Learning; Bioinformatics

1. Introduction

Genomic research in the context of the current and future development has brought more insights into the genetic factors that underlie complex diseases, especially those of neurological nature. This increase in knowledge is mainly due to the combination of high-throughput sequencing techniques with advanced computational approaches, which give the investigators a vast view of the molecular mechanisms of neurological diseases. However, these issues have not been resolved and the task of making sense of raw DNA sequences remains as difficult as ever, presenting a great opportunity for developing new approaches to unlock the potential of the vast amounts of genetic data available.

E-mail address: cs_vasavi@blr.amrita.edu

In response, our work proposes a novel framework that incorporates DNA sequence embeddings to help identify brain diseases at an early stage. These embeddings learn to map genomic DNA sequences into real-valued vectors while preserving relevant structural and functional characteristics. Our method utilizes the state-of-art deep learning architectures such as Convolutional and Artificial Neural Networks to extract the features from these embeddings. These features make them highly useful in diagnosing neurological disorders ranging from Alzheimer's, Parkinson's and various neurodevelopmental disorders.

The main contribution of the study is:

- Development of a novel framework for genetic brain disease classification using DNA sequence embeddings and advanced machine learning techniques, particularly a custom Stacking Classifier model.
- Achieving high accuracy (94.8%) in classifying five major neurological disorders, outperforming individual models including CNNs.
- Demonstrating the effectiveness of integrating DNA BERT embeddings to capture relevant genetic information for disease classification.

This paper is organized as follows: Section I introduces the concept of genomic research in neurological diseases and the potential of DNA sequence embeddings for disease classification. Section II provides a detailed literature review on existing approaches for DNA sequence classification, brain disease prediction, and machine learning applications in genomics. Section III discusses the proposed methodology, including data collection, DNA BERT embeddings generation, dataset curation, and the development of various machine learning models, including the custom Stacking Classifier. Section IV presents the results of the implementations and a comparative analysis between different approaches. Section V interprets the results, discussing the implications of the model performances and the effectiveness of the DNA embedding approach. Section VI concludes the study, summarizing the key findings and their potential impact on genetic disease classification and personalized medicine. Finally, Section VII outlines future enhancements, suggesting potential avenues for improving the models and expanding the research scope.

2. Literature Survey

The paper by Ramalakshmi et al. (2021)[1] presents a literature survey on the classification of DNA sequences using deep learning techniques such as CNNs, hybrid models combining CNNs with LSTM and bidirectional LSTM networks. The authors review various existing studies employing deep learning models for DNA sequence classification across different applications, including identifying pathogens, detecting virus effects, and drug design. They discuss the use of techniques like CNNs, DNNs, N-gram probabilistic models, extreme gradient boosting algorithms with hybrid features, feature selection and stacking methods, alignment-free models based on K-mer feature extraction, and linear classifiers. The survey also covers the application of deep learning for predicting specific viruses like SARS-CoV-2, improving code construction for DNA sequences, classifying complex sequences with ensemble decision tree approaches, and classifying DNA sequences for cancer patients.

The paper by A. Raza et al. (2022)[2] analyzes a dataset with 44 attributes, including patient information, family history, medical test results, and genetic information. The proposed ETRF feature engineering approach combines the ET and RF algorithms to extract features from the genomes dataset, which are then used as input for learning techniques to predict genetic disorders and their types. The study introduces the ETRF method for feature extraction, which is utilized to develop learning models that predict genetic disorders. The research by S. Victor[3] focuses on the importance of prediction in medical diagnosis in the field of bioinformatics, highlighting its role in improving healthcare efficiency and reducing complexities in treatment. The use of DNA sequences and structure information is proposed as a means to get the accuracy and optimization of disease. The study by Ismaeel et al. (2013) proposes a novel method for predicting disease based on mutations in the gene sequence, specifically focusing on breast cancer mutations. The method utilizes bioinformatics techniques such as FASTA and CLUSTALW to detect malignant mutations that increase the probability of cancer, and the backpropagation algorithm is trained to classify whether a patient has the disease or not[3].

The paper by Ismaeel et al. (2013)[4] on a machine learning model is to predict the disease due to mutations in the gene sequence and the researchers have applied the model especially for the breast cancer mutations. The method

utilizes bioinformatics techniques such as FASTA and CLUSTALW to detect malignant mutations that increase the probability of cancer. Cross entropy is trained with expected malignant mutations of particular genes like BRCA1, BRCA 2 genes in breast cancer and used for the prediction of whether the patient is affected or not. Cancer diagnosis is the crucial step of cancer treatment and the paper stresses on the significance of early and accurate diagnosis and opens the possibilities provided by computer-aided diagnosis. Further researches will be directed with the goal of creating a regional data base of genetic diseases and creating an integrated system of early diagnostic for patients with genetic diseases using the identified method. The paper also mentions the use of feedforward back-propagation neural networks for classifying malignant mutations for breast cancer. Classification of skin disease based on color and texture feature is the main concern presented in the paper under consideration [5] by K. V. Swamy. It employs the hue, saturation and value color model to transform the RGB color space to other more comprehensible space and infer the current state of affairs regarding texture based feature extraction techniques for skin disease identification. The algorithm proposed in the paper is, DT and SVM, and the classification is based on entropy, variance, and maximum histogram value of the feature, namely, HSV, and the performance analysis was done based on accuracy.

This paper by Pandey(2023) et al introduces a solution for early disease identification by leveraging DNA sequence classification, particularly crucial in the context of fast-spreading viruses like COVID-19. Utilizing samples from NCBI's Genbank, the proposed framework matches patient DNA samples to identify diseases, aided by a new hot vector-based representation for feature extraction. Through extensive experimentation and comparison with traditional classifiers, including CNN, SVM, KNN, Decision Trees, and RNN, the proposed method achieves a high accuracy rate of 93.9 percent, demonstrating its potential in enhancing public health efforts through early disease prediction and management[6].The paper by Senol et al (2011) addresses the complexity of disease prediction by considering the interplay of multiple genomes and their impact on disease susceptibility. By leveraging DNA sequences and Bayesian network pathway analysis, the system aims to determine the probabilistic levels of disease occurrence based on mutations in causal and associated genes. Emphasizing the importance of genetic information in disease prediction, the paper discusses methodologies and architectures to effectively identify disease markers and pathways. Case studies on diseases like Type-1 Diabetes and Crohn's disease further validate the system's potential in enhancing disease prevention strategies.[7]

In this study, Wu et al. [8] selected six frequently occurring facial skin diseases which are acne, freckles, rosacea, senile spots, seborrheic keratosis, and skin prolapse, and they compared five different networks, ResNet-50, Inception-v3, DenseNet121, Xception, and Inception-ResNet-v2 using the largest clinical image dataset of skin diseases in China Alshahrani also emphasized that Inception-ResNet-v2 was the best one which downloaded to a definite understanding that deep learning was promising in medical images processing. In the same vein, Ahmad et al. [9] introduced a discriminative feature leaning approach to hyper fine-tune ResNet152 as well as InceptionResNet-V2 model using a triplet loss function that boosts the efficiency of skin disease image classification. One of the remarkable alternatives of theirs is based on deep CNNs to embed input images into Euclidean space and distinguishes the images employing L-2 distance among images learning discriminative features yielding a comparatively high accuracy compared to many advanced methods. This study utilized a dataset of human face skin disease images from a hospital in Wuhan, China, to show the applicability of the conceived framework in outcomping conventional methods.

Saied et al. [10] employed a new approach with the S-parameter of six antennas placed around the head to monitor the shifts in the dielectric properties of the brain that are characteristic of AD. In their study, they employed differing methodologies that assist with machine learning, such as Logistic Regression, demonstrated a 98. For example, in distinguishing between Alzheimer's disease (AD) stages, it achieved 97% accuracy, outperforming the traditional MRI and PET scans. Gunduz [11] developed two deep learning frameworks that were based on CNNs for classifying PD using several sets of vocal features. The source of data for the study was the UCI Machine Learning Repository, and the author found marked enhancement in classification accuracy and discriminative ability over conventional techniques. Nair et al. [12] developed another study in adopting an ensemble approach employed feature selection and classification method that employs SVM, KNN, and Decision Tree algorithm that improves performance by using a majority voting system. Veetil et al. [13] have reviewed five major architectures of deep learning and the authors showed while there exists overlap, VGG19 performs best at classifying the T2-Weighted MRI scans That are from PPMI dataset and reiterated the usefulness of transfer learning in medical image classification.

The paper by Sarada Jayan et al. (2022)[16] compares different machine learning techniques for cancer classification by employing RNA-seq data, considering 5 types of cancerous tumors are breast carcinoma, KIRC refers to

kidney renal carcinoma, and LUAD is an acronym for lung adenocarcinoma as does PRAD for prostate adenocarcinoma while the last is COAD for colon adenocarcinoma. Carried out using exploratory analysis, PCA is subsequently used for data feature reduction; the paper examines six algorithms: Naïve Bayes, knn , logistic regression, decision tree, random forest, and svm. The overall working and performance analysis indicates that out of all the algorithms, SVM and random forest have surpassed all with accuracy, SVM has achieved an optimum accuracy of 100% and its time complexity is lesser than that of Random Forest, AUC-ROC, and class separability is as well higher in SVM that in Random Forest. A recent paper by Kiran Kumar et al. (2019)[17] offer a review of the various machine learning techniques utilized in early cancer predicting and prognosis models for the various types of cancer such as breast, oral, skin, colon, and lung cancer among others. Regarding computational methods and methodologies of the past and present for machine learning in cancer prediction and diagnosis, it describes the advantages, disadvantages, and critical points related to existing methodologies with which one may need help developing new machine learning methodologies for greater disease-specific prediction and diagnosis.

P. Jyothi et al. [20] put forward an approach using the Support Vector Machine (SVM) in connection with DNAPred to predict hereditary diseases by analyzing the DNAs sequences inherited from parents to children. The base of the methodology is to improve the accuracy of the forecasts using the neural networks, and to emphasize that the selected method – SVM – showed the best result among all the classifiers that were tested. Many studies have used human genome data with an attempt to diagnose severe dengue prognoses using various machine learning techniques; the C. Davi et al. [18] study used SNP selection support vector machine and an artificial neural network for classification. A traditional strategy they used bearing evidence of high accuracy, sensitivity, and specificity was pointing towards the fact that the genetic context offers the possibility of defining the phenotype in dengue. Dolci et al. [19] suggest a deep MMG architecture incorporating f/ sMRI & retiring/learnable genomes for AD. functional MRI, structural MRI, retiring genomes, learnable genomes In an attempt to fill this gap, their model leverages knowledge transferred from generative adversarial networks in an effort to perform better in predicting admittance to the Alzheimer's even when presented with incompletely acquired data.

3. Methodology

In this section, we detail the development and implementation of our dataset,Machine Learning & deep Learning Algorithms and our custom stacking classifier model. This section outlines the systematic approach taken to evaluate and integrate multiple predictive algorithms into a cohesive and optimized framework, ensuring both high accuracy and robust performance in our predictive tasks.

3.1. Data Collection

The methodology for retrieving data from the NCBI Nucleotide database uses a focused web scraping approach with the BioPython library Entrez module, which allows programmatic access to biological data. Specific queries are formulated for diseases such as Alzheimer's Disease, Amyotrophic Lateral Sclerosis, Down Syndrome, Epilepsy, and Parkinson's Disease, using the format "[Disease Name][Title] AND Homo sapiens[Organism]" to ensure searches are confined to human-related studies and exclude non-human data. The Entrez.esearch function conducts searches within the 'nucleotide' database, retrieving a predefined number of entries (typically 100) that match the disease-specific criteria, producing a list of sequence identifiers. These identifiers are then used with the Entrez.efetch function to download the sequence data in FASTA format. Sequences are processed using SeqIO.parse for parsing and extracting both the sequence IDs and their corresponding sequences, facilitating systematic data extraction and subsequent analysis tailored to specific research needs.

3.2. BERT Embeddings

We utilized the 'bert-base-uncased' BERT model to create embeddings from nucleotide sequences associated with various genetic brain diseases. This model is adept at capturing deep contextual relationships within data, which

is crucial for understanding complex biological sequences. We transformed these sequences into 6-mer tokens to align with BERT's input requirements, ensuring that each token adequately represents the genetic information. After tokenizing, these sequences were fed into the BERT model, and we extracted embeddings by averaging the outputs of the last hidden state, resulting in a 768-dimensional vector for each sequence. The final dataset, structured into a DataFrame, includes columns for the sequence ID, its associated disease label, and the 768 embedding dimensions.

3.3. Pseudocode:

Approach taken for generating the sequence embeddings from NCBI is described in Algorithm I

Algorithm 1 Generate Sequence Embeddings from NCBI Database

Require: BERT tokenizer and model initialized
Ensure: Sequence embeddings are generated and saved

```
Input: disease_name
Initiate search for nucleotide data associated with the disease
Retrieve sequence IDs from NCBI database
if sequence IDs are found then
    Fetch sequences using sequence IDs
    if sequences are successfully retrieved then
        Generate embeddings using BERT model
        Prepare a data structure to store sequence IDs, embeddings, and disease name
        Save the data structure to an Excel file
        Print confirmation of saved embeddings
    else
        Print error message: No sequences retrieved
    end if
else
    Print error message: No sequences found
end if
```

3.4. Dataset Details

In our study, we have compiled a dataset comprising 5,000 genetic sequences sourced from the NCBI database. To maintain class balance, we included 1000 sequences from each of five different diseases: Alzheimer's Disease, Amyotrophic Lateral Sclerosis, Down Syndrome, Epilepsy, and Parkinson's Disease. Each genomic sequence was transformed into embeddings using DNA BERT, resulting in 768-dimensional vectors that encapsulate the features of the sequences. Consequently, our final dataset dimensions are 1000 rows, each corresponding to a sequence, and 768 columns, representing the feature vectors derived from DNA BERT embeddings.

3.5. Dataset Quality Enhancements

To ensure the integrity and non-redundancy of our dataset, we implemented a rigorous two-fold verification system. This approach was necessitated by the conversion of all sequences into embedding vectors, enabling more nuanced similarity analyses. Given the importance of avoiding redundancy and duplicates in our data, we employed multiple methods to achieve this.

First, since the sequences were transformed into 783-dimensional embedding vectors, we utilized cosine similarity to identify and eliminate redundant entries. However, in the field of biology, even a single letter difference can signify a critical mutation, providing essential insights into diseases. Therefore, we implemented a second layer of verification using the Needleman-Wunsch algorithm, a global alignment technique. This additional step allowed us to thoroughly purify the data, enhancing its overall quality and reliability.

3.5.1. Cosine Similarity Analysis

We used cosine similarity to compare the embedding vectors of our 5000 sequences, resulting in 12,497,500 pairwise comparisons. Our analysis revealed:

- 10% of comparisons resulted in 100% similarity
- 60% of comparisons showed similarity above 95%

These findings indicated significant potential redundancy within our dataset, necessitating further investigation and refinement.

3.5.2. Needleman-Wunsch Algorithm for Secondary Verification

To address sequences with high similarity (≥95% but <100%), we implemented a secondary verification using the Needleman-Wunsch algorithm. We randomly selected 100 sequence pairs from those with ≥95% cosine similarity for this analysis.

Our two-fold verification system allowed us to

- Remove truly redundant data with high confidence
- Retain sequences that represented unique biological entities despite high embedding similarity

3.5.3. Dataset Refinement

Following these analyses, we refined our dataset by removing 100% similar sequences, eliminating near-identical sequences confirmed by Needleman-Wunsch alignment, and flagging ambiguous cases for expert review. The resulting refined dataset served as the basis for our subsequent analyses, with a portion reserved for testing purposes.

This comprehensive approach to dataset curation combined efficient computational methods with intensive biological sequence analysis, balancing computational efficiency with biological relevance. The result was a dataset of significantly improved quality and reliability for our genetic disease classification of brain disorders.

3.6. Model Training

3.6.1. Machine Learning& Deep Learning Models:

As part of our work, we used different machine learning techniques, such as Random Forest, XGBoost, K-NN, SVM, CatBoost, AdaBoost, Decision Tree, Naive Bayes, Convolutional Neural Network (CNN), and Multi-Layer Perceptron (MLP). All these algorithms have been selected for their ability to address different aspects of the genetic data that goes into the classification of diseases.

3.6.2. Hyper-parameter Tuning

To find the best set of hyperparameters for different machine learning models that will enhance the predictive accuracy of disease classification from genetic data, we used the Grid Search technique. Grid Search is an iterative process that consists of comparing several combinations of hyperparameters which have been predetermined and then using cross-validation to check their performance. Because of this, we can seamlessly cover all possible combinations of the parameters and determine the best values for each model. The values from the grid search which are seen in the reference table, Table I include parameters such as the number of estimators, depth of the trees, learning rates and other model specific parameters for models like Random Forest, XGBoost, SVM and many more. These values reflect the settings we used in this study that yielded the highest-performing models on our data,.

3.6.3. Custom Model

Building on the foundational research of previous studies, we have developed a novel custom stacking classifier model, the mechanics of which are detailed in the flow chart below (see Fig1). This model operates by initially evaluating a range of algorithms to identify those that deliver the most accurate predictions. The selected algorithms are then utilized as base models, with the highest performing algorithm serving as the meta-model. This strategic configuration results in a custom model that is uniquely tailored to optimize performance. The primary advantage of

Table 1. Hyperparameter Table

Algorithm	**Final Hyperparameter Values**
Random Forest	n_estimators: 100 criterion: 'gini' max_features: 'auto' max_depth: 30 min_samples_split: 2 min_samples_leaf: 1 bootstrap: True
XGBoost	n_estimators: 100 max_depth: 6 learning_rate: 0.1 subsample: 0.9 colsample_bytree: 0.9
K-Nearest Neighbors	n_neighbors: 5 weights: 'uniform' p: 2 leaf_size: 10
Support Vector Machine	C: 1 kernel: 'rbf' gamma: 'scale'
CatBoost	depth: 6 learning_rate: 0.01 iterations: 100
AdaBoost	n_estimators: 100 learning_rate: 0.1
Decision Tree	max_depth: 20 min_samples_split: 2 min_samples_leaf: 1
Naive Bayes	No hyperparameters to tune
Convolutional Neural Network	filters: 64 kernel_size: 3 Dense: 128 Dropout: 0.5
Multi-Layer Perceptron	hidden_layer_sizes: (100,) alpha: 0.0001 solver: 'adam'

our custom stacking classifier is its enhanced predictive accuracy and robustness, achieved by effectively combining the strengths of individual models into a cohesive system. This approach, which represents a novel contribution to the field, ensures that our model not only advances theoretical understanding but also enhances practical applications in predictive analytics.

As shown in the pseudo code below we have taken

4. Results & Discussion

Our research aimed to address two primary questions:

Q1: How effective are DNA sequence embeddings in classifying genetic brain diseases?

Algorithm 2 Stacking Classifier Approach

Require: Dataset D, base models set M
Ensure: Trained stacking classifier

```
Split D into D_train and D_test
for each model m_i ∈ M do
    Evaluate m_i on D_train via cross-validation
    if performance of m_i > threshold then
        Add m_i to selected models S
    end if
end for
Train each model in S on D_train
Generate predictions P from S on D_train
Train meta-model M_meta on P
Generate final predictions using S and M_meta on D_test
return Trained stacking classifier
```

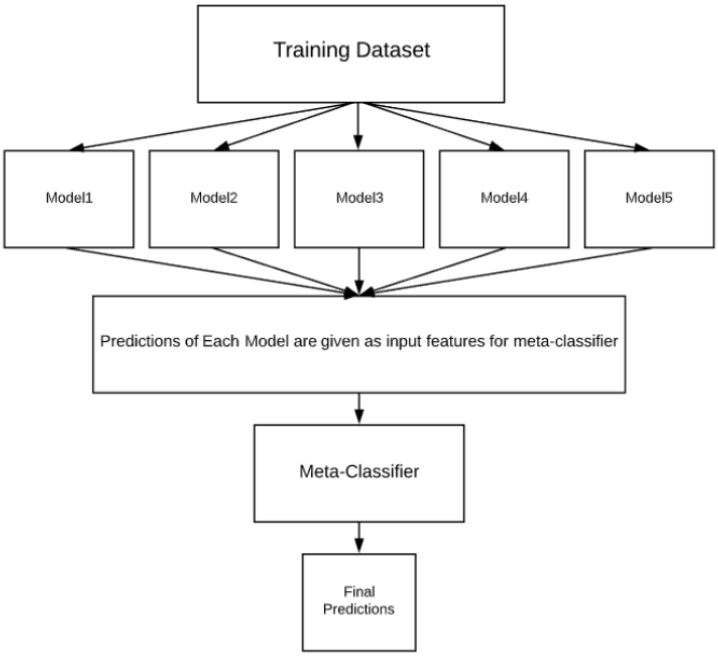

Fig. 1. Flow diagram of the custom Stacking Classifier model architecture

A1: The DNA BERT embeddings proved highly effective, enabling our models to achieve high accuracy in classifying five major neurological disorders. The custom Stacking Classifier achieved 94.8% accuracy, demonstrating the power of these embeddings in capturing relevant genetic information.

Q2: How does the performance of our custom Stacking Classifier compare to traditional machine learning and deep learning models?

A2: The custom Stacking Classifier outperformed individual models, including sophisticated deep learning techniques like CNNs. It achieved 94.8% accuracy compared to CNN's 93%, showcasing the potential of ensemble methods in handling complex genomic data. In this comparison (as seen from table II), we see a diverse set of algorithms, including traditional methods like Random Forest and SVM, as well as more advanced techniques like XGBoost and CNN (Convolutional Neural Network). Our Stacking Classifier emerges as the top performer with an impressive accuracy of 0.948, narrowly surpassing the CNN's 0.93. This result is particularly noteworthy as CNNs are often

considered state-of-the-art for many classification tasks, especially in image recognition. The Stacking Classifier's success likely stems from its ability to combine multiple models, leveraging the strengths of different algorithms to make more robust predictions. It's like having a panel of experts, each contributing their unique insights to reach a final decision.

Table 2. Model Performance on DNA BERT-Embeddings

Algorithm	**Accuracy**
Random Forest	0.8700
XGBoost	0.9250
KNN	0.8757
SVM	0.8770
CatBoost	0.4900
AdaBoost	0.5300
Decision Tree	0.7295
Naive Bayes	0.6200
CNN	0.9300
MLP	0.7400
Stacking Classifier	0.9480
Voting Classifier	0.9100

In this comparison(as seen from table II), we see a diverse set of algorithms, including traditional methods like Random Forest and SVM, as well as more advanced techniques like XGBoost and CNN (Convolutional Neural Network). Our Stacking Classifier emerges as the top performer with an impressive accuracy of 0.948, narrowly surpassing the CNN's 0.93. This result is particularly noteworthy as CNNs are often considered state-of-the-art for many classification tasks, especially in image recognition. The Stacking Classifier's success likely stems from its ability to combine multiple models, leveraging the strengths of different algorithms to make more robust predictions. It's like having a panel of experts, each contributing their unique insights to reach a final decision. The high performance of both the Stacking Classifier and CNN underscores the power of ensemble methods and deep learning in modern machine learning. While the CNN excels at automatically learning hierarchical features from data, the Stacking Classifier's strength lies in its adaptability and ability to capture diverse patterns. This comparison highlights that sometimes, a well-designed ensemble can match or even outperform sophisticated neural networks, demonstrating that there's often more than one path to achieving excellent results in machine learning tasks. As with can see with the Fig 3.., The blue

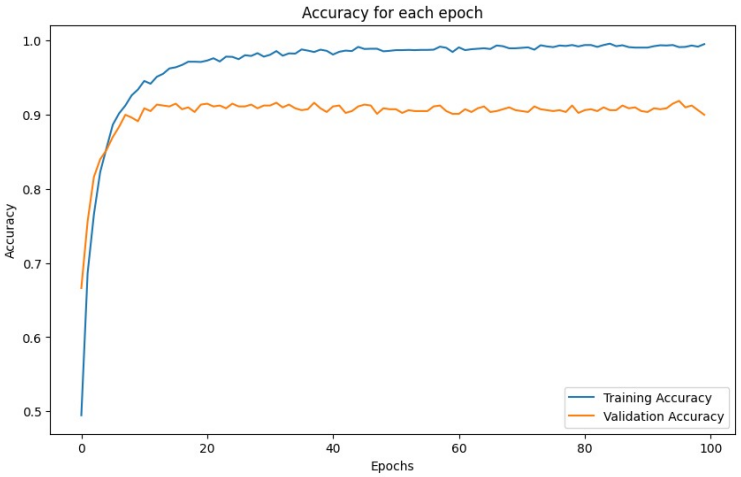

Fig. 2. CNN model accuracy over training epochs

line represents training accuracy, while the orange line shows validation accuracy. Both start low but quickly improve

in the first few epochs, indicating rapid initial learning. The training accuracy continues to climb steadily, eventually reaching nearly 100%. This suggests the model is learning the training data extremely well. However, the validation accuracy plateaus around 90-92% after the initial rapid improvement. This gap between training and validation accuracy is a classic sign of overfitting. The model is memorizing specific patterns in the training data that don't generalize well to new, unseen data (represented by the validation set). Despite the overfitting, the model still achieves strong validation performance.

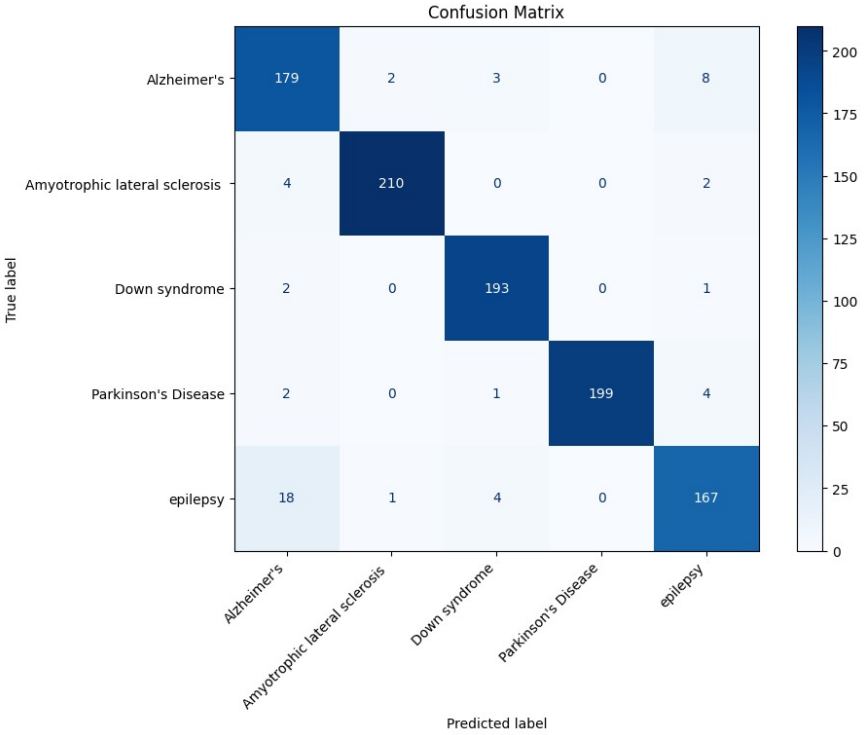

Fig. 3. Confusion matrix for neurological disorder classification

The confusion matrix(Fig 3.) provides insights into the performance of a classification model for five neurological conditions: Alzheimer's, Amyotrophic lateral sclerosis (ALS), Down syndrome, Parkinson's Disease, and epilepsy. The diagonal elements represent correct classifications, showing strong performance across all categories. Alzheimer's (179 correct), ALS (210 correct), Down syndrome (193 correct), Parkinson's Disease (199 correct), and epilepsy (167 correct) all have high accurate prediction rates. However, there are some misclassifications worth noting. The model sometimes confuses Alzheimer's with other conditions, particularly epilepsy (8 cases). Epilepsy also shows the most misclassifications, with 18 cases mistaken for Alzheimer's. This suggests that distinguishing between Alzheimer's and epilepsy might be challenging for the model, possibly due to some overlapping symptoms or features in the data. The few misclassifications indicate areas where the model could potentially be improved, particularly in differentiating between Alzheimer's and epilepsy.

The bar chart in Fig 4 compares the performance of the Stacking Classifier and the CNN model across three key metrics: Precision, Recall, and F1 Score. The Stacking Classifier outperforms the CNN in both Precision and F1 Score, registering values of 0.80 and 0.78 respectively, against the CNN's 0.76 and 0.73. This indicates that the Stacking Classifier not only predicts more relevant results (as reflected in the higher Precision) but also maintains a better balance between Precision and Recall, as evidenced by the superior F1 Score.

However, CNN achieves a slightly higher recall of 0.72 compared to the stacking classifier, 0.70, suggesting that CNN is marginally better at identifying all relevant cases within the dataset. The enhanced performance of the Stacking Classifier in the other two metrics can be attributed to its combination of multiple base models, which allows it to

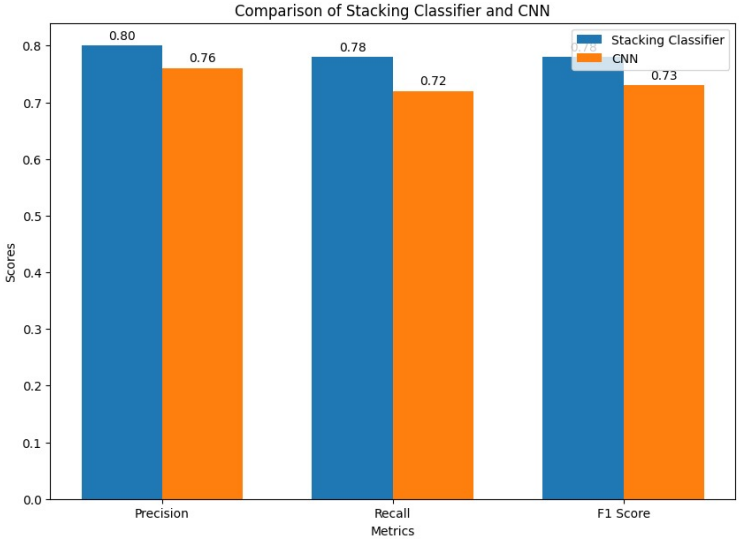

Fig. 4. Performance comparison of Stacking Classifier vs CNN

capture more diverse patterns within the data, ultimately resulting in a more robust and accurate prediction model overall.

5. Conclusion

Our study introduces a novel approach to genetic brain disease classification using DNA sequence embeddings and advanced machine learning techniques. The custom Stacking Classifier model achieved an impressive 94.8% accuracy, outperforming individual models including CNNs (93% accuracy) in classifying five major neurological disorders. This demonstrates the power of ensemble methods in handling complex genomic data, rivaling sophisticated deep learning techniques. The integration of DNA BERT embeddings proved highly effective in capturing relevant genetic information, enabling our models to discern subtle patterns associated with different brain disorders. This research opens new avenues for applying machine learning in genomics and personalized medicine, potentially revolutionizing diagnostic procedures and contributing to a deeper understanding of the genetic basis of neurological disorders.

6. Future Enhancements

As the next steps, there are some possibilities to refine the models in the future. First, trying out other possibilities of base models to be stacked in the Stacking Classifier could reveal other combinations that can improve the performance of the algorithm. Further, using more data or using data augmentation techniques could help the models in having a wider range of examples to understand better. Finally, the future work could include the application of enhanced regularization approaches and the investigation of more recent architectures for CNN, which could further enhance the model's generalization ability and yield better performance in future applications.

References

[1] Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C., & Suresh Gnana Dhas, C. (2021). Analysis of DNA sequence classification using CNN and hybrid models. Computational and Mathematical Methods in Medicine, 2021.

[2] Raza, A., Rustam, F., Siddiqui, H. U. R., Diez, I. D. L. T., Garcia-Zapirain, B., Lee, E., & Ashraf, I. (2022). Predicting genetic disorder and types of disorder using chain classifier approach. Genes, 14(1), 71.

[3] Victor, S.P., OPTIMIZED PREDICTION IN MEDICAL DIAGNOSIS USING DNA SEQUENCES AND STRUCTURE INFORMATION.

[4] Ismaeel, A. G., & Ablahad, A. A. (2013). Novel method for mutational disease prediction using bioinformatics techniques and backpropagation algorithm. arXiv preprint arXiv:1303.0539.
[5] Swamy KV, Divya B. Skin disease classification using machine learning algorithms. In2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4) 2021 Dec 16 (pp. 1-5). IEEE.
[6] Mathur, G., Pandey, A., & Goyal, S. (2023). A comprehensive tool for rapid and accurate prediction of disease using DNA sequence classifier. Journal of Ambient Intelligence and Humanized Computing, 14(10), 13869-13885.
[7] Kotiang, S., & Eslami, A. (2020). A probabilistic graphical model for system-wide analysis of gene regulatory networks. Bioinformatics, 36(10), 3192-3199.
[8] Z. Wu et al., "Studies on Different CNN Algorithms for Face Skin Disease Classification Based on Clinical Images," in IEEE Access, vol. 7, pp. 66505-66511, 2019, doi: 10.1109/ACCESS.2019.2918221.
[9] B. Ahmad, M. Usama, C. -M. Huang, K. Hwang, M. S. Hossain and G. Muhammad, "Discriminative Feature Learning for Skin Disease Classification Using Deep Convolutional Neural Network," in IEEE Access, vol. 8, pp. 39025-39033, 2020, doi: 10.1109/ACCESS.2020.2975198.
[10] I. M. Saied, T. Arslan and S. Chandran, "Classification of Alzheimer's Disease Using RF Signals and Machine Learning," in IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology, vol. 6, no. 1, pp. 77-85, March 2022, doi: 10.1109/JERM.2021.3096172.
[11] H. Gunduz, "Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets," in IEEE Access, vol. 7, pp. 115540-115551, 2019, doi: 10.1109/ACCESS.2019.2936564.
[12] A. J. Nair, R. Rasheed, K. Maheeshma, L. Aiswarya and K. R. Kavitha, "An Ensemble-Based Feature Selection and Classification of Gene Expression using Support Vector Machine, K-Nearest Neighbor, Decision Tree," 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2019, pp. 1618-1623, doi: 10.1109/ICCES45898.2019.9002041.
[13] I. K. Veetil, E. A. Gopalakrishnan, V. Sowmya and K. P. Soman, "Parkinson's Disease Classification from Magnetic Resonance Images (MRI) using Deep Transfer Learned Convolutional Neural Networks," 2021 IEEE 18th India Council International Conference (INDICON), Guwahati, India, 2021, pp. 1-6, doi: 10.1109/INDICON52576.2021.9691745.
[14] P. Lehotay-Kéry, G. Kicska and A. Kiss, "Genome classification with deep learning using heuristic algorithms for hyper-parameter optimization," 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA), Hammamet, Tunisia, 2023, pp. 1-6, doi: 10.1109/INISTA59065.2023.10310599.
[15] Ning, K., Chen, B., Sun, F., Hobel, Z., Zhao, L., Matloff, W & Alzheimer's Disease Neuroimaging Initiative. (2018). Classifying Alzheimer's disease with brain imaging and genetic data using a neural network framework. Neurobiology of aging, 68, 151-158.
[16] A. A, R. R, R. Paul, S. Jayan, A. Thakur and N. P. Tv, "Comparative Study of Cancer Classification by Analysis of RNA-seq Gene Expression Levels," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-7, doi: 10.1109/ICCCNT54827.2022.9984600.
[17] M Kiran Kumar, Divya Udayan J,2019, " A Study on Machine Learning Techniques in Cancer Disease Prediction and Diagnosis", Indian Journal of Public Health Research & Development, Vol 10, Issue 4, pp. 157-162.(SCOPUS Index)
[18] C. Davi et al., "Severe Dengue Prognosis Using Human Genome Data and Machine Learning," in IEEE Transactions on Biomedical Engineering, vol. 66, no. 10, pp. 2861-2868, Oct. 2019, doi: 10.1109/TBME.2019.2897285.
[19] G. Dolci et al., "A deep generative multimodal imaging genomics framework for Alzheimer's disease prediction," 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 2022, pp. 41-44, doi: 10.1109/BIBE55377.2022.00017.
[20] P. Jyothi and P. Ajitha, "Computerized Prediction of Hereditary Diseases Through DNA Sequence Using Support Vector Machine (SVM)," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2021, pp. 953-958, doi: 10.1109/ICECA52323.2021.9675867.