Deep Learning Analyses as a Tool to Diagnose Rare Diseases Based on Symptoms: A Comparative Study of Different Methodologies

Bezalel Itzhaky*, Ezra Ella, Ph.D*.

affiliation: *Afeka Academic College

Keywords: *rare diseases, deep learning, PCA, neural networks, dataset analysis, Decision Tree, ML, AI, medical informatics*

Abstract

In this work, we applied an advanced deep learning analytical method to train a model for disease prediction based on symptoms, using a rare disease dataset from Kaggle. The dataset contains disease names and corresponding symptoms experienced by each patient. There are 773 unique diseases and 377 symptoms, with approximately 246,000 rows of data. We observed that for very rare diseases—where the number of occurrences is sometimes fewer than five—machine learning (ML) and deep learning (DL) methods can produce highly unstable models due to overfitting or poor representation of certain categories in the test group. We also noted that removing samples with insufficient representation led to a more stable model, albeit with significantly lower accuracy. Finally, we tested the top 22 diseases (each having at least 1000 observations) and found that we could train a stable model with good accuracy.

Introduction

Rare disease patients suffer from the rarity of their conditions in two major ways. First, treatments for rare diseases are not always accessible, given their low prevalence and the high cost of developing new therapies. Second, the scarcity of cases often leads to misdiagnosis, since many doctors are unfamiliar with these diseases and because false positives in the general population are more common than true positives.

In this work, we obtained a rare disease dataset from Kaggle comprising 773 unique diseases and 377 symptoms, with approximately 246,000 rows. This dataset was artificially generated, preserving symptom severity and disease occurrence probability. Several distinct symptom groups can indicate the same disease, and in some rows, only a single symptom may be present. Such cases suggest a high correlation between that symptom and the disease in question. A larger number of rows for a given disease corresponds to a higher real-world probability of occurrence.

For some rare diseases where an effective treatment exists, timely diagnosis is critical: an undiagnosed disease can significantly worsen health outcomes due to a lack of proper treatment. Our goal in this study is to investigate the feasibility of using deep learning methods to predict rare diseases based on symptoms, as well as to assess the impact of data preprocessing, feature selection, and different observation counts on model accuracy. We employ both traditional machine learning and advanced neural network models and compare their performances under various data manipulation strategies, including principal component analysis (PCA).

Literature Review

The application of deep learning (DL) and machine learning (ML) techniques to diagnose rare diseases based on symptom data is a rapidly evolving area of medical informatics. Traditional diagnostic methods often struggle with rare diseases due to their low prevalence, which leads to limited clinical experience and data availability (McCarthy, 2019). Consequently, patients with rare diseases frequently experience diagnostic delays and misdiagnoses, highlighting the urgent need for innovative diagnostic tools (Tafuri, 2025).

Recent studies have demonstrated the potential of ML and DL in analyzing complex medical datasets to improve diagnostic accuracy. For instance, support vector machines (SVMs) and decision trees have been employed to identify patterns in symptom data and predict disease occurrence (Hossain, 2023). However, these methods often face challenges with high-dimensional data and imbalanced datasets, which are common in rare disease studies (Castro-Espin, 2022).

Neural networks, particularly deep learning architectures, have shown promise in overcoming these limitations. Deep learning models can automatically learn hierarchical representations of features from raw data, enabling them to capture intricate relationships between symptoms and diseases (LeCun et al., 2015). Studies have explored various neural network architectures, including fully connected networks and convolutional neural networks (CNNs), for disease prediction using electronic health records and symptom data (Rajkomar et al., 2019).

Dimensionality reduction techniques, such as principal component analysis (PCA), are frequently used to mitigate the curse of dimensionality and improve model performance (Jolliffe & Cadima, 2016. Jolliffe, 2016). PCA can transform high-dimensional symptom data into a lower-dimensional space while preserving the most significant variance, thereby enhancing the efficiency and accuracy of diagnostic models.

Despite these advancements, several gaps remain in the literature. Many studies focus on common diseases, and the application of DL to rare diseases is still limited. Furthermore, the impact of dataset size and class imbalance on model performance in rare disease diagnosis requires further investigation. The current study aims to address these gaps by systematically evaluating the performance of various ML and DL models on a large, artificially generated rare disease dataset, focusing on the effects of data preprocessing, feature selection, and model architecture.

Methodology

Dataset Selection and Overview

We obtained from Kaggle a rare disease and symptoms dataset containing 773 unique diseases and 377 symptoms, with around 246,000 rows. The dataset was artificially generated, preserving symptom severity and disease occurrence probability. Several distinct groups of symptoms can be indicators of the same disease, and in some rows, only a single symptom may be present, indicating a high correlation with that disease. A larger number of rows for a given disease corresponds to its higher real-world probability of occurrence. Similarly, if only one symptom is present in a row, that symptom likely has a stronger correlation with the disease than any single symptom in a multi-symptom row. All symptoms in the dataset are binary (0 or 1).

To test the effect of data preprocessing on model quality, we applied three different data treatment approaches: **Minimal Intervention**: Removing categories with fewer than five incidences. **Moderate Intervention**: Additional data processing, ensuring a minimal representation of each class in both the training and test sets. **Aggressive Intervention**: Selecting the top 22 diseases, each with more than 1000 occurrences.

Removing Low-Represented Groups to Allow Model Fitting

Because some diseases had very few observations, we progressively removed low-representation diseases from the dataset. First, we removed diseases with fewer than two observations, which excluded only 20 samples but did not resolve model-fitting issues. Next, we removed diseases with fewer than five observations, eliminating an additional 99 samples and finally allowing the model to train properly.

Setting Validation and Test Sets

We typically used 70–80% of the dataset for training and 20–30% for validation and testing. The data always had 378 columns for symptoms, though the number of rows varied according to the filtering approach. For minimal data processing, for example, we used: **Training set**: ~132,739 rows, 378 columns. **Validation set**: ~28,444 rows, 378 columns. **Test set**: ~28,445 rows, 378 columns.

Using a Cost-Sensitive Random Forest Classifier

We employed a cost-sensitive Random Forest classifier with adjusted parameters to handle the challenges posed by some classes having very few samples. Traditional ML models often struggle with imbalanced datasets, where underrepresented classes lead to biased predictions. Cost-sensitive learning addresses this issue by assigning higher misclassification penalties to rare classes, encouraging the model to prioritize their accurate classification.

Fully Connected Neural Network Architecture

We implemented a fully connected neural network using the Sequential API from Keras. The architecture consisted of an input layer, two hidden layers, and an output layer: **First hidden layer**: 64 neurons, ReLU activation, followed by a Dropout layer (rate = 0.2). **Second hidden layer**: 32 neurons, ReLU activation. **Output layer**: Softmax activation for multi-class classification

We compiled the model using the Adam optimizer with a sparse categorical cross-entropy loss, suitable for multi-class classification with integer-encoded labels. We used early stopping (patience of 5 epochs) to prevent overfitting and selected the best-performing model weights. Training spanned up to 50 epochs with a batch size of 32, and final performance was measured on an independent test set.

Hyperparameter Tuning for Neural Network

We conducted a grid search over learning rates (0.001, 0.01, 0.1), batch sizes (16, 32, 64), and epochs (30, 50, 70). Each model used the same architecture (64 neurons + dropout, 32 neurons, softmax output) and was evaluated on a balanced training set (stratified) , with early stopping based on max Recall The best-performing combination of hyperparameters was selected based on the highest test accuracy.

Dataset Preprocessing and Feature Selection

A comprehensive preprocessing and feature selection pipeline was used to enhance data quality and reduce dimensionality: A. **Data Acquisition and Duplication**: The dataset was imported and duplicated to preserve integrity. B. **Exploratory Data Analysis (EDA)**: We examined the distribution of selected symptoms (e.g., anxiety, nervousness, depression, shortness of breath) via KDE plots and histograms. Descriptive statistics were computed to detect skewness, outliers, and data imbalance. C. **Frequency-Based Feature Selection**: We retained only symptoms present in at least 5% of the rows, eliminating low-frequency variables. This approach improved model generalization and computational efficiency. D. **Final Dataset Composition**: We retained the most frequently observed symptom variables and the disease target column. Debugging checks ensured the integrity of the filtered dataset.

Performance Degradation Simulation and Model Training
To test model robustness under data loss, we randomly removed 50% of the dataset. After preprocessing (including SMOTE for class imbalance), we trained a cost-sensitive Random Forest classifier on the reduced data and evaluated it on a test set. The classification report was used to quantify performance degradation.

Enhanced Neural Network Architecture
We developed an enhanced neural network with additional layers and batch normalization: **Input layer. Dense layers**: 128, 64, and 32 neurons with ReLU activation. **Batch normalization and 30% dropout** after the first two hidden layers. **Output layer**: Softmax activation.

The Adam optimizer and categorical cross-entropy loss were used, along with early stopping (patience = 5). This enhanced network showed improved accuracy compared to baseline models.

Dimensionality Reduction
We applied Principal Component Analysis (PCA) to reduce dimensionality. Using n_components=0.95 retained 95% of the variance in the original dataset. The PCA transformation was fitted on the training data and applied to both validation and test sets. We then compared the performance of a Decision Tree classifier trained on PCA-reduced data with one trained on the original data.

Data Analysis, Code Writing, and Text Corrections
The code for data preprocessing, normalization, and model construction (both ML and neural network) was initially generated using ChatGPT (OpenAI) and Gemini 2.0 (Google), then polished and modified by the authors to fit specific requirements. The article text was also reviewed using ChatGPT for English language refinement. Additional methodological details can be found in the appendix.

Results

Number of Observations Ranged from 0 to ~1200; Binary Symptom Representation

After loading the dataset, we generated a statistical and visual report to understand its structure and determine necessary modifications. We observed that the data was artificially produced and exhibited a step-like format (Figure 1a). The number of observations per disease ranged from 0 to around 1200. All symptoms were binary (0 or 1), and there was unexplained variation in symptom patterns for the same disease (Figure 1c). After aggressive data filtering, we ended up with the top 22 diseases, each having the largest number of observations (Figure 1d).

Poor Performance for Minimal and Moderate Data Manipulation with Cost-Sensitive Random Forest

Both minimal and moderate data manipulation approaches yielded low F1 scores and high support values (Table 1). This result suggests the model struggled to accurately predict less common classes. After aggressive data manipulation (retaining only diseases with high observation counts), the initial results were poor (F1 score = 0, data not shown), but once we employed a stratified split, we obtained higher accuracy, recall, precision, and F1 scores (Figure 2).

Insufficient Representation of Groups in the Test Set Leads to Faulty Models

Comparing minimal and moderate preprocessing, we found that minimal preprocessing sometimes produced extremely high accuracy (~99.9%), whereas moderate preprocessing produced accuracy around 72–73% (Table 2). We attribute the minimal preprocessing result to an artifact of overfitting and poor representation of certain classes in the test set. For the aggressive data filtering, both methods yielded similar results, likely because each disease had enough observations for stable model training.

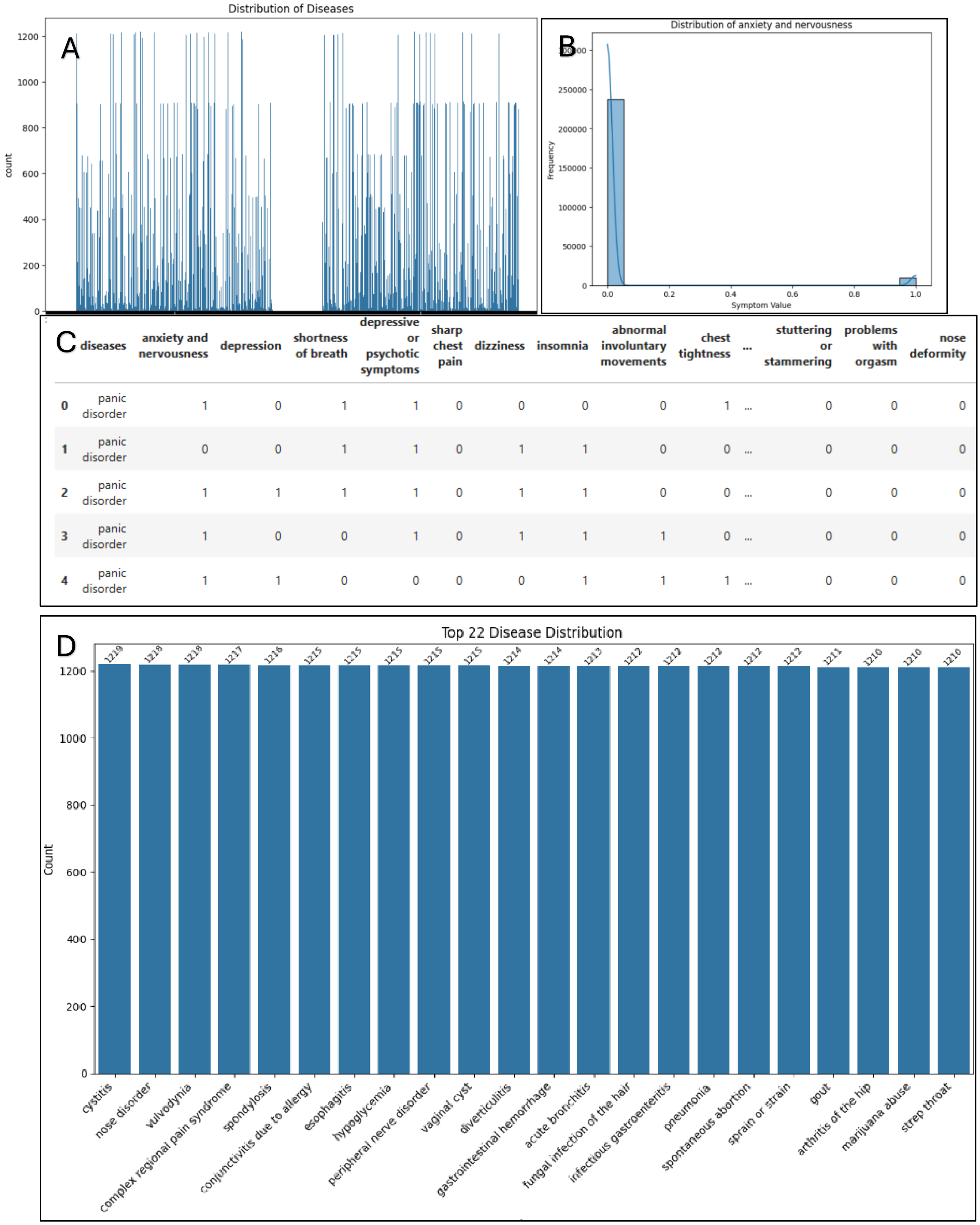
Fully Connected Neural Network Also Affected by Imbalanced Data

Using the fully connected neural network described in the methodology, we observed similar test accuracies for both minimal and moderate data processing (Table 3). However, minimal preprocessing required roughly twice as many epochs to converge, indicating the network struggled with noisy, imbalanced data. Notably, the neural network's test accuracy for minimal preprocessing dropped from 99% (with the Random Forest) to 92%, whereas moderate preprocessing improved from 72% to about 81.5%. Area Under the Curve (AUC) graphs (Figure 3a,b) indicated overfitting in both minimal and moderate data approaches, as evidenced by a significant gap between training and validation AUC. For aggressive data processing (Figure 3c), results were similar to the Decision Tree, with a very high train and validation AUC from the first epochs, which reflect high accuracy and significantly lower loss—about ten times smaller compared to the minimal and moderate processing models, indicating a more stable and accurate model.

Data Quality and Learning Rate Have the Strongest Effect on Accuracy

We tuned hyperparameters (learning rate: 0.001, 0.01, 0.1; batch sizes: 16, 32, 64; epochs: 30, 50, 70) and observed that learning rate had the largest impact on accuracy. At a learning rate of 0.001, batch size and epoch variations had only a minor effect. However, at 0.01, batch size influenced test accuracy more strongly. A learning rate of 0.1 consistently produced poor results.

Under aggressive data processing, accuracy remained high even at a learning rate of 0.01, and batch size or number of epochs had minimal impact. These findings highlight the superior importance of data quality and class representation over hyperparameter adjustments.



	diseases	anxiety and nervousness	depression	shortness of breath	depressive or psychotic symptoms	sharp chest pain	dizziness	insomnia	abnormal involuntary movements	chest tightness	...	stuttering or stammering	problems with orgasm	nose deformity
0	panic disorder	1	0	1	1	0	0	0	0	1	...	0	0	0
1	panic disorder	0	0	1	1	0	1	1	0	0	...	0	0	0
2	panic disorder	1	1	1	1	0	1	1	0	0	...	0	0	0
3	panic disorder	1	0	0	1	0	1	1	1	0	...	0	0	0
4	panic disorder	1	1	0	0	0	0	1	1	1	...	0	0	0

Figure 1 Data distribution and structure. (a) In this image, each line on the X-axis represents a rare disease, and the height of the line on the Y-axis corresponds to the number of observations in the analyzed dataset. It is evident that some diseases have few or zero observations, while others have up to 1200 observations per disease. The abrupt step-jumps (e.g., no observations between ~1200 and 900 or between ~900 and 700) indicate that the dataset was artificially generated. (b) This panel shows a representative image of the binary distribution of symptom data. Most observed cases (approximately 230,000) are negative for the specific symptom “Anxiety and nervousness” (tagged as 0), while about 6,000 observations are positive (tagged as 1). (c) This table illustrates the structure of the data: for

each rare disease, every one of the 378 symptom columns has a binary value (0 or 1). Notice that the same disease can have different tag values across observations, implying internal noise and variability within the groups.(d) The top 22 rare diseases were selected based on the number of observations (count).

Minimal intervention				Moderate intervention				Aggressive intervention			
precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support
		0.54	37926			0.47	37914			0.95	4006
0.59	0.54	0.49		0.55	0.51	0.45		0.95	0.95	0.95	
0.84	0.54	0.62		0.79	0.47	0.55		0.95	0.95	0.95	

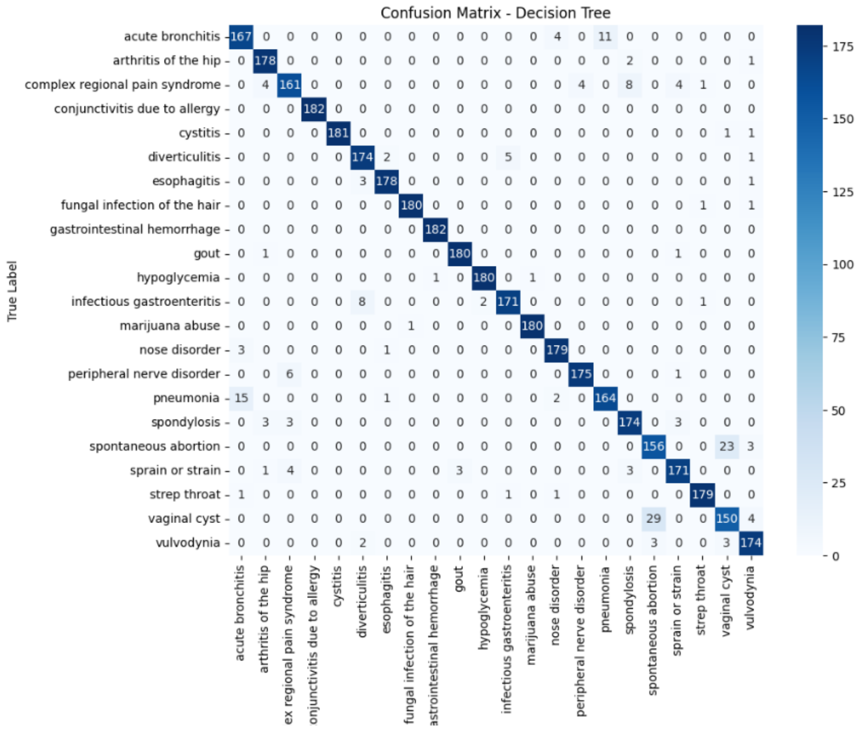
Table 1 Random forest results: Cost-sensitive Random Forest analysis was performed under three conditions: minimal, moderate, and aggressive data processing. For both minimal and moderate processing, the F1 score—as well as precision and recall—are very low despite extremely high support values. This is a classical case where the model struggles with common classes and fails to provide accurate predictions.

Intervention	Accuracy	Precision	Recall	F1-score
Minimal	0.9993	0.9992	0.9993	0.9991
Moderate	0.7233	0.7368	0.7233	0.7258
Aggressive	0.9526	0.9527	0.9526	0.9525

Table 2Desicion tree: Decision tree performance scores. The results indicate that using data with minimal processing provides better performance compared to moderate processing, with significant differences in accuracy, recall, and F1 score. Although the aggressive intervention produced the highest score values, the identical values across all metrics may suggest an issue with data balancing.

Intervention	Epochs	Accuracy	Loss	Val accuracy	Val Loss	Test Accuracy
Minimal	31	0.885	0.3371	0.9249	0.2258	0.9246
Moderate	15	0.8848	0.3741	0.8286	0.5652	0.8152
Aggressive	25	0.9687	0.0628	0.9621	0.0654	0.9618

Table 3 Neuronal network with hyperparameters scores: Neural network performance scores with hyperparameter tuning. For each intervention level, the same neural network architecture was trained. The epoch values reflect the point at which further training did not lead to significant improvements (i.e., model convergence). Minimal and moderate processing required relatively high epoch counts and exhibited higher loss values compared to the aggressive method, indicating a higher level of noise in these models.



	acute bronchitis	arthritis of the hip	ex regional pain syndrome	onjunctivitis due to allergy	cystitis	diverticulitis	esophagitis	fungal infection of the hair	astrointestinal hemorrhage	gout	hypoglycemia	infectious gastroenteritis	marijuana abuse	nose disorder	peripheral nerve disorder	pneumonia	spondylosis	spontaneous abortion	sprain or strain	strep throat	vaginal cyst	vulvodynia
acute bronchitis	167	0	0	0	0	0	0	0	0	0	0	0	0	4	0	11	0	0	0	0	0	0
arthritis of the hip	0	178	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1
complex regional pain syndrome	0	4	161	0	0	0	0	0	0	0	0	0	0	0	4	0	8	0	4	1	0	0
conjunctivitis due to allergy	0	0	0	182	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cystitis	0	0	0	0	181	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
diverticulitis	0	0	0	0	0	174	2	0	0	0	0	5	0	0	0	0	0	0	0	0	0	1
esophagitis	0	0	0	0	0	3	178	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
fungal infection of the hair	0	0	0	0	0	0	0	180	0	0	0	0	0	0	0	0	0	0	0	1	0	1
gastrointestinal hemorrhage	0	0	0	0	0	0	0	0	182	0	0	0	0	0	0	0	0	0	0	0	0	0
gout	0	1	0	0	0	0	0	0	0	180	0	0	0	0	0	0	0	0	1	0	0	0
hypoglycemia	0	0	0	0	0	0	0	0	1	0	180	0	1	0	0	0	0	0	0	0	0	0
infectious gastroenteritis	0	0	0	0	0	8	0	0	0	0	2	171	0	0	0	0	0	0	0	1	0	0
marijuana abuse	0	0	0	0	0	0	0	1	0	0	0	0	180	0	0	0	0	0	0	0	0	0
nose disorder	3	0	0	0	0	0	1	0	0	0	0	0	0	179	0	0	0	0	0	0	0	0
peripheral nerve disorder	0	0	6	0	0	0	0	0	0	0	0	0	0	0	175	0	0	0	1	0	0	0
pneumonia	15	0	0	0	0	0	1	0	0	0	0	0	0	2	0	164	0	0	0	0	0	0
spondylosis	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	174	0	3	0	0	0
spontaneous abortion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	156	0	0	23	3
sprain or strain	0	1	4	0	0	0	0	0	0	3	0	0	0	0	0	0	3	0	171	0	0	0
strep throat	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	179	0	0
vaginal cyst	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	150	4
vulvodynia	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	3	0	0	3	174

Figure 2 Confusion matrix: Confusion matrix for the top 22 rare diseases with the highest number of observations. The close similarity between true and predicted labels for most cases is reflected in high scores for accuracy, recall, precision, and F1.

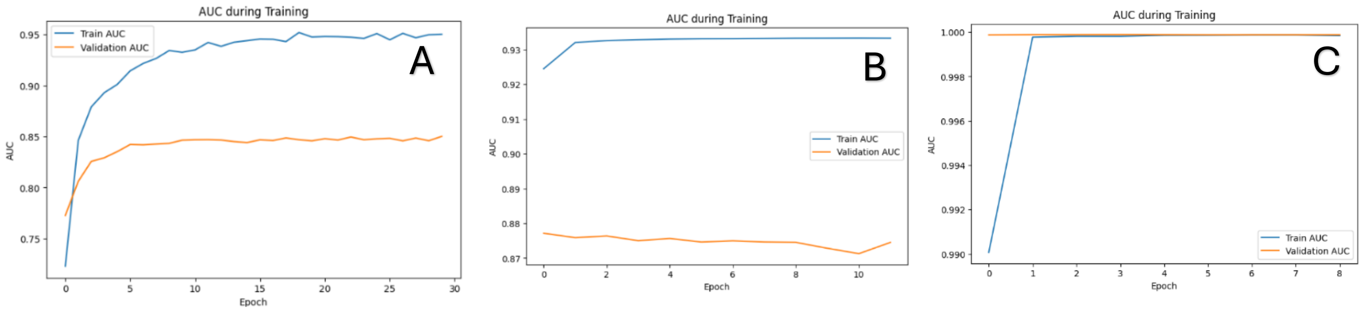


Figure 3 Train and validation AUC: AUC curves for minimal and moderate data processing suggest overfitting. In both cases, the training AUC (blue line) rapidly climbs and remains above 0.9, while the validation AUC (orange line) shows a clear gap, indicating that the model is overfitting to the training data. (a) minimal data processing, (b) moderate data processing and (c) Aggressive data processing.

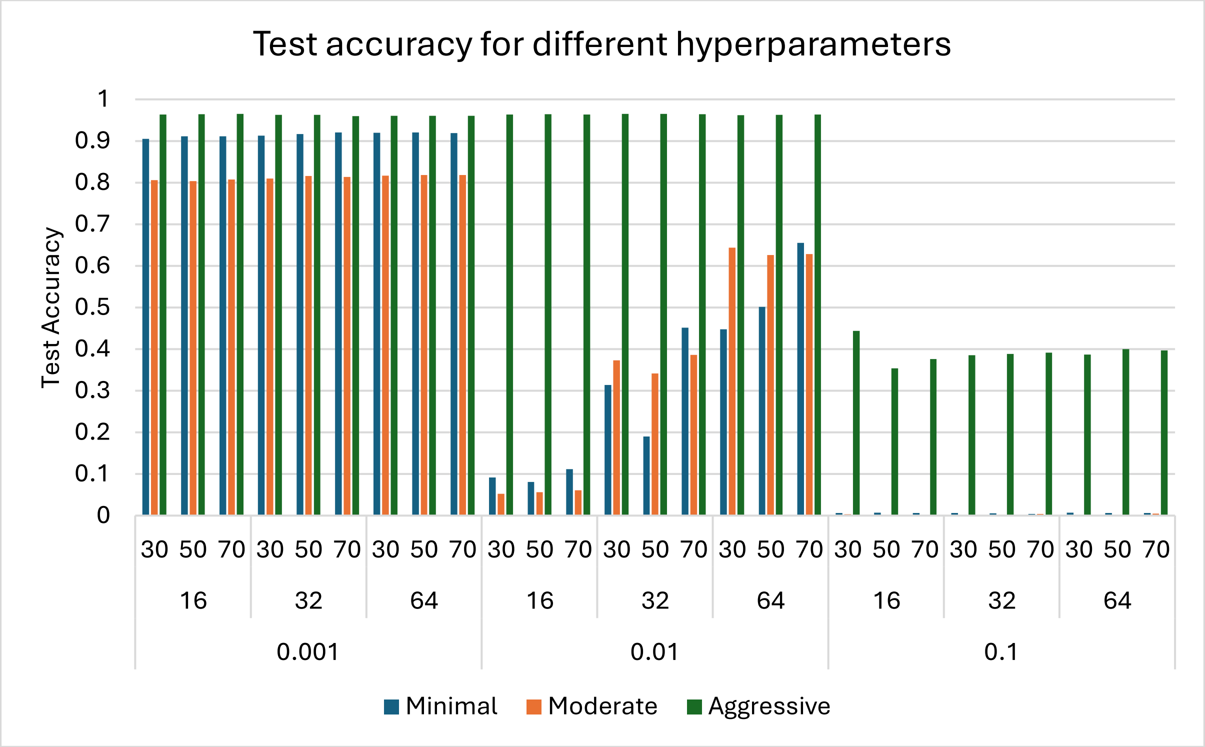


Figure 4 Test accuracy for each method and different hyperparameters: Test accuracy for different hyperparameter configurations. The Y-axis represents test accuracy, while the X-axis shows the various conditions based on epochs, batch size, and learning rate. The graph clearly indicates that the best test accuracy was achieved with a small learning rate of 0.001. Additionally, for a learning rate of 0.01, there is a positive correlation between batch size, the number of epochs, and accuracy.

Data Degradation Reveals the Model Struggles with Rare Disease Prediction

When we randomly removed 50% of the data, the training set shrank to ~98,777 rows and the test set to ~24,695 rows. Under both minimal and moderate data processing, the model predominantly predicted the "False" category (~99.86% of cases), failing almost entirely to detect the "True" category (Table 4). This phenomenon is common for rare diseases, as models default to the majority class to maximize overall accuracy.

High Performance in the Confusion Matrix Can Be Misleading for Rare Diseases

The confusion matrix shows high overall accuracy due to the model's correct classification of the dominant negative (not diseased) class. However, the near-zero accuracy in detecting positive (rare disease) cases reveals the model's failure to generalize for rare conditions.

The effect of model enhancement is neglected in comparison to data quality

We conducted training sessions on the minimal and moderate processing datasets using a neural network that was twice as dense and deep (see Table 5). Surprisingly, the enhanced neural network did not improve the final model's performance; we were unable to achieve the same results as those obtained using aggressive data processing with the standard neural network. This finding underscores the critical importance of data quality over model complexity in determining final model performance.

PCA Analysis Shows the Necessity of Accurate Data Preprocessing

For the PCA analysis, minimal data preprocessing yielded a “striped” PCA plot with a single dominant principal component explaining virtually all variance (Figure 5a). In contrast, moderate preprocessing (Figure 5b) produced a more balanced distribution of variance across multiple principal components. Since all features are binary, better representation of classes across the dataset allows PCA to capture the real structure of the data rather than being skewed by a few dominant features. When using aggressive data processing (Figure 5c) the dimension of the principle component was dramatically reduced, due to smaller data set, and also by the ability of the model to group observation in unique clusters, providing an image with visibly distinct clusters.

Balancing of data is critical for understanding variance

We plotted the variance for each model to compare the effects of different data processing methods. The primary difference between minimal and moderate data processing was in data balancing, indicating that the dimensions and content of the datasets were very similar. However, the PCA variance plots reveal notable differences. For the minimal processing (Figure 6a), the entire variance is explained by a single component, suggesting that the data structure is dominated by one feature or group of features. In contrast, the moderate processing (Figure 6b) distributes the variance across many components, with each component explaining only a small portion of the total variance. This highlights the importance of data balancing in capturing a more nuanced and true representation of the data. For aggressive data processing (Figure 6c), the total explained variance was relatively low, which is consistent with our observation that the model trained on aggressively processed data exhibited low overall variance.

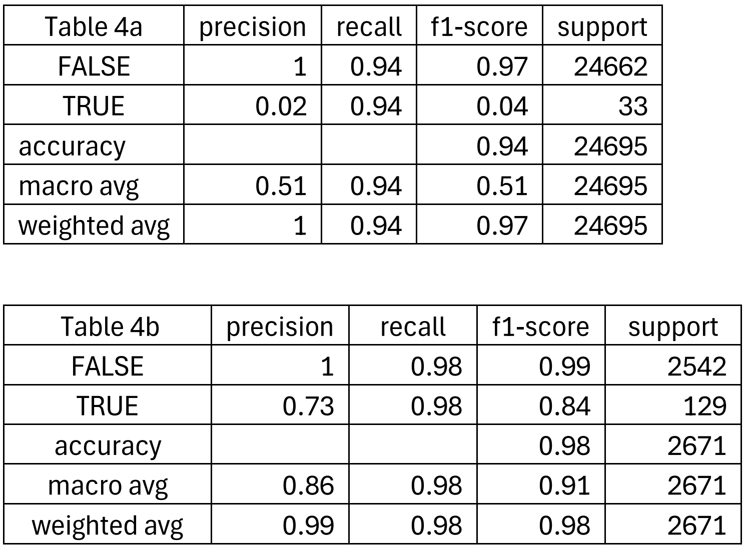


Table 4a	precision	recall	f1-score	support
FALSE	1	0.94	0.97	24662
TRUE	0.02	0.94	0.04	33
accuracy			0.94	24695
macro avg	0.51	0.94	0.51	24695
weighted avg	1	0.94	0.97	24695

Table 4b	precision	recall	f1-score	support
FALSE	1	0.98	0.99	2542
TRUE	0.73	0.98	0.84	129
accuracy			0.98	2671
macro avg	0.86	0.98	0.91	2671
weighted avg	0.99	0.98	0.98	2671

Figure 4 Results of Confusion matrix for degraded dataset: Confusion matrix for the degraded dataset using aggressive processing. After data degradation, both minimal and moderate data processing (a) accurately detect healthy cases (False) but perform poorly in identifying positive cases. In contrast, aggressive data processing (b) shows significant improvement, with an F1-score of 0.84. This demonstrates that data quality (with a sufficient number of observations per case) can be more critical than data quantity.

Intervention	Epochs	Accuracy	Loss	Val accuracy	Val Loss	Test Accuracy
Minimal	20	0.8541	0.4583	0.9266	0.2284	0.92385
Moderate	19	0.8722	0.4217	0.8352	0.5196	0.8235

Table 5 Results for neuronal network architecture enhancement: The results of enhanced neuronal network for the same dataset but with a difference level of data processing. Although we made the neuronal network about twice deeper and larger, there were no differences in the results of accuracy or loss, suggesting that the noise in the data can't be compensated by deeper network. For the aggressive data processing method there was no need to run this deeper learning process, as the results were already excellent, showing the importance of high quality dataset for deep learning models.

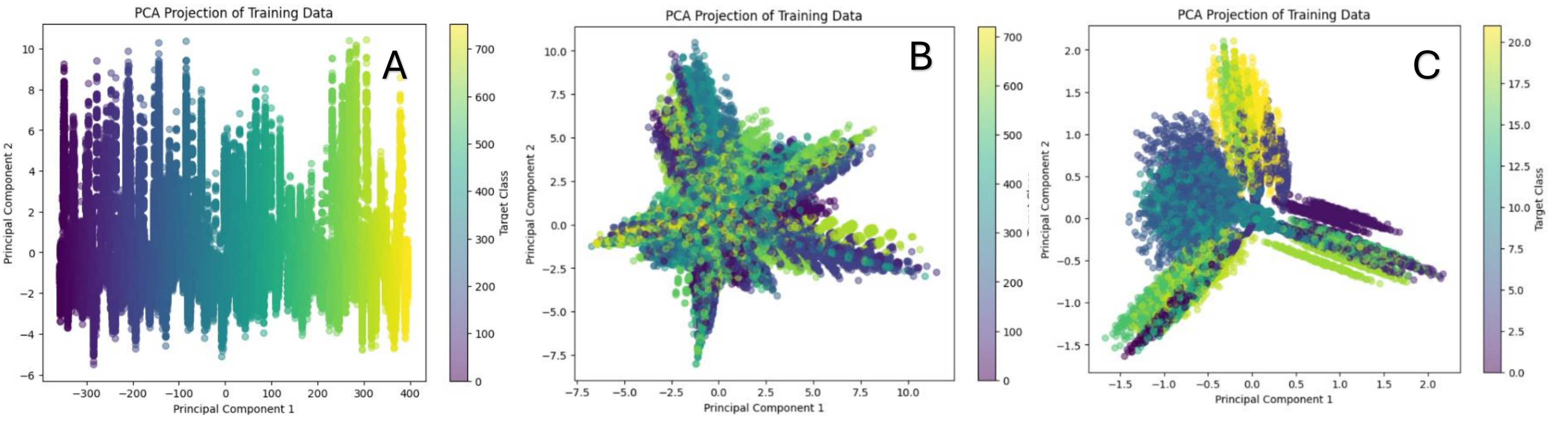


Figure 5 : (a) For minimal data processing, the PCA analysis shows a high range for component 1 (approximately -300 to 400), suggesting issues with dimensionality reduction. (b) For moderate data processing, the range of component 1 is significantly reduced (from around 700 to about 18), and clearer group patterns emerge, indicating that groups sharing similar characteristics are better separated. (c) For aggressive data processing, the dimensions are reduced further (approximately 4×4 compared to the initial 700×16 for minimal processing), allowing for clear visual separation of groups based on color and distribution.

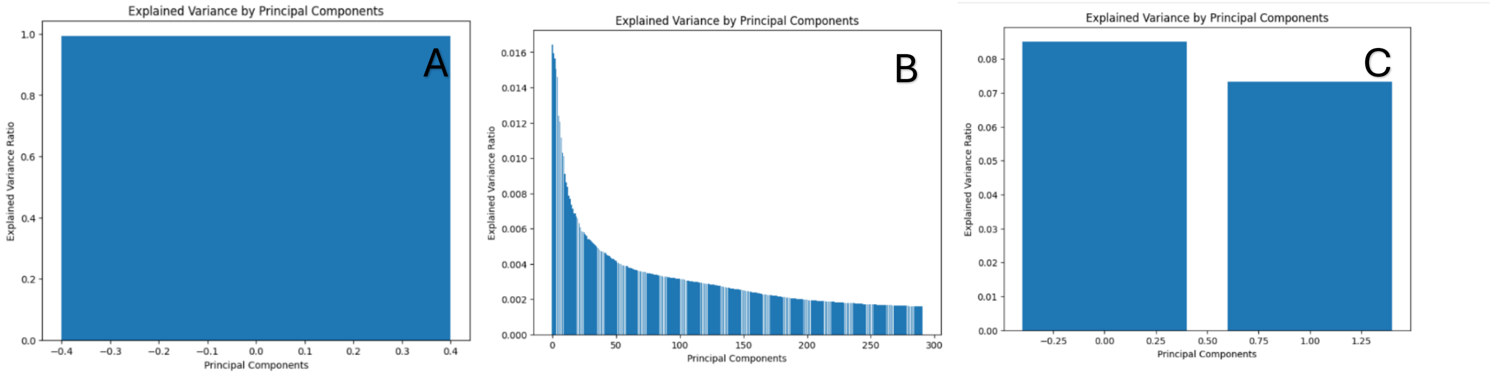


Figure 6 explained variance by PCA: (a) Explained variance by PCA for minimally processed data: This plot shows a single, large bar spanning the entire range of principal components, with an explained variance ratio of 1.0. (b) Explained variance by PCA for moderately processed data: The plot exhibits a more typical PCA result, where the explained variance ratio decreases as the number of principal components increases; the first few components explain a significant portion of the variance, with subsequent components contributing progressively less. (c) Explained variance by PCA for aggressively processed data: Most of the variance is explained by just two components, likely due to the relatively low variance and noise in the dataset after aggressive processing.

Discussion and Conclusions

We aimed to detect rare diseases based on binary symptom data using various ML and DL methods. Our results consistently showed that **data quality and class representation** are more critical than hyperparameter tuning alone. Although we obtained high accuracy scores under minimal data manipulation, deeper analysis revealed significant overfitting and poor generalization—particularly problematic for rare disease detection.

Accuracy alone is not sufficient for evaluating model performance in rare disease contexts. Metrics such as F1 score, precision, recall, and loss values are essential to reveal whether the model is genuinely learning or merely exploiting the imbalance in the data.

When we adopted **aggressive data filtering** (retaining only diseases with ample observations), both ML and DL models converged more quickly, demonstrated smaller loss values, and avoided overfitting. By contrast, minimal and moderate preprocessing led to noisy, imbalanced data that hampered training stability and reliability.

PCA analysis further highlighted how imbalanced classes can distort dimensionality reduction. A large proportion of variance captured by a single principal component often indicates skewed data, while a more typical PCA curve with multiple components suggests balanced feature representation.

Additional Insight of ChatGPT:
Future research could explore **ensemble approaches** (e.g., combining Random Forest with neural networks) and **data augmentation** (beyond SMOTE) tailored to rare diseases. Generative models or transfer learning might help address extreme data imbalance by synthesizing realistic minority-class samples or leveraging knowledge from related diseases. Collaboration with domain experts is also vital, as medical expertise can guide feature engineering, symptom weighting, and the validation of synthetic data to ensure that models capture clinically meaningful patterns.

Future Work

In this study, we highlighted the effect of data quality on model performance for rare disease prediction. The dataset used was synthetic, and the group sizes varied considerably. The main purpose of this Kaggle dataset was to serve as a benchmark for testing different deep learning methods in rare disease diagnosis. As a next step, it would be valuable to test the best-performing architectures on **real-world datasets** and further refine model strategies for small sample sizes. Investigating **transfer learning** or advanced **bootstrap methods** in such low-data scenarios could be particularly promising. Additionally, exploring how synthetic data can best be aligned with true clinical distributions would help ensure that performance gains translate into real-world diagnostic improvements.

References

1. Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P. D., Pisani, A. R., & Turner, K. (2023). Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-Making: A Systematic Review. Computers in Biology and Medicine, 155, 106649.
2. Castro-Espin, C., & Agudo, A. (2022). The Role of Diet in Prognosis among Cancer Survivors: A Systematic Review and Meta-Analysis of Dietary Patterns and Diet Interventions. Nutrients, 14(2), 348.
3. McCarthy, C., Kokosi, M., & Bonella, F. (2019). Shaping the Future of an Ultra-Rare Disease: Unmet Needs in the Diagnosis and Treatment of Pulmonary Alveolar Proteinosis. Current Opinion in Pulmonary Medicine, 25(5), 450–458.
4. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436–444.
5. Rajkomar, A., Oren, E., Chen, K. T., Dai, A. M., Hajaj, N., Hardt, M., et al. (2019). Scalable and Accurate Deep Learning with Electronic Health Records. NPJ Digital Medicine, 2(1), 1–10.
6. Shire. (2013). Rare Diseases: Challenges and Opportunities in the European Union. Shire Human Genetic Therapies.
7. Tafuri, G., Bracco, A., & Grueger, J. (2025). Access and Pricing of Medicines for Patients with Rare Diseases in the European Union: An Industry Perspective. A 2025 Update. Expert Review of Pharmacoeconomics & Outcomes Research. Advance online publication.
8. Jolliffe, I. T., & Cadima, J. (2016). Principal Component Analysis: A Review and Recent Developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.